Max-Affine Regression via first-order methods*

Seonho Kim† and Kiryung Lee†

Abstract. We consider regression of a max-affine model that produces a piecewise linear model by combining affine models via the max function. The max-affine model ubiquitously arises in applications in signal processing and statistics including multiclass classification, auction problems, and convex regression. It also generalizes phase retrieval and learning rectifier linear unit activation functions. We present a non-asymptotic convergence analysis of gradient descent (GD) and mini-batch stochastic gradient descent (SGD) for max-affine regression when the model is observed at random locations following the sub-Gaussianity and an anti-concentration with additive sub-Gaussian noise. Under these assumptions, a suitably initialized GD and SGD converge linearly to a neighborhood of the ground truth specified by the corresponding error bound. We provide numerical results that corroborate the theoretical findings. Importantly, SGD not only converges faster in run time with fewer observations than alternating minimization and GD in the noiseless scenario but also outperforms them in low-sample scenarios with noise.

Key words. Max-affine regression, gradient descent, stochastic gradient descent, non-convex optimization.

AMS subject classifications. 90C26

1. Introduction.

The *max-affine* model combines k affine models in the form of

$$y = \max_{j \in [k]} \left(\langle \boldsymbol{x}, \boldsymbol{\theta}_j^\star \rangle + b_j^\star \right) \tag{1.1}$$

to produce a piecewise-linear mutivariate functions, where $\boldsymbol{x}$ and y respectively denote the covariate and the response, and $[k]$ denotes the set $\{1, \ldots, k\}$. The max-affine model frequently arises in applications of statistics, machine learning, economics, and signal processing. For example, the max-affine model has been adopted for multiclass classification problems [7, 9] and simple auction problems [31, 34].

We consider a regression of the max-affine model in (1.1) via least squares

$$\min_{\{\boldsymbol{\theta}_j, b_j\}_{j=1}^k} \frac{1}{2n} \sum_{i=1}^n \left(y_i - \max_{j \in [k]} (\langle \boldsymbol{x}_i, \boldsymbol{\theta}_j \rangle + b_j) \right)^2 \tag{1.2}$$

from statistical observations $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ potentially corrupted with noise. A suite of numerical methods has been proposed to solve the nonconvex optimization in (1.2) (e.g., [30, 42, 19, 1]). The *least-squares partition algorithm* [30] iteratively refines the parameter estimate by alternating between the partition and the least-squares steps when the number of affine models k is known a priori. The partitioning step classifies the inputs $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ with respect to the maximizing affine models given estimated model parameters. The least-squares step updates the parameters for each affine model by using the corresponding observations. Later variations of the alternating minimization algorithm used an adaptive search for unknown k [19, 1].

*Submitted to the editors on March 17, 2024.

Funding: This work was supported in part by NSF CAREER Award CCF 19-43201.

†The Ohio State University, Columbus, Ohio (kim.7604@osu.edu, kiryung@ece.osu.edu).

The consistency of these estimators has been derived. In more recent papers, Ghosh et al. [12, 13, 14] established finite-sample analysis of the *alternating minimization* (AM) estimator [30] for the special case when the observations are generated from a ground-truth model. One can interpret their analysis through the lens of the popular *teacher-student framework* [29]. This framework has been widely adopted in statistical mechanics [29, 10] and machine learning [49, 15, 48, 22]. It provides a theoretical understanding of how a specific model is trained and generalized through a ground-truth generative model [22]. In this framework, a max-affine model (student) is trained by data generated from a ground-truth max-affine model (teacher) from k fixed affine models. By using the provided data, the student model recovers parameters that produce the ground-truth model via AM. Since the max affine model is invariant under the permutation of the component affine models, the minimizer to (1.2) is determined only up to the corresponding equivalence class. Ghosh et al. [14] established a finite-sample analysis of AM under the standard Gaussian covariate assumption with independent stochastic noise. They showed that a suitably initialized alternating minimization converges linearly to a consistent estimate of the ground-truth parameters along with a non-asymptotic error bound. Moreover, they proposed and analyzed a spectral method that provides the desired initialization. They also further extended the theory to a generalized scenario with relaxed assumptions on the covariate model [12, 13].

In this paper, we present analogous theoretical and numerical results on max-affine regression by first-order methods including *gradient descent* (GD) and *stochastic gradient descent* (SGD). The first-order methods have been widely used to solve various nonlinear least squares problems in machine learning [16, 11, 39, 24]. We observe that GD and SGD also perform competitively on max-affine regression compared to AM. In particular, SGD converges significantly faster (in run time) than AM in a noise-free scenario. Figure 1 compares AM, GD, and a mini-batch SGD on random 50 trials of max-affine regression where the ground-truth parameter vectors $\{\boldsymbol{\beta}_j^\star\}_{j=1}^k$ are selected randomly from the unit sphere. Covariates are independently generated from either $\mathrm{Normal}(\mathbf{0}, \boldsymbol{I}_{500})$ or $\mathrm{Unif}[-\sqrt{3}, \sqrt{3}]^{\otimes 500}$. We plot the median of relative errors versus the average run time where the relative error is calculated as

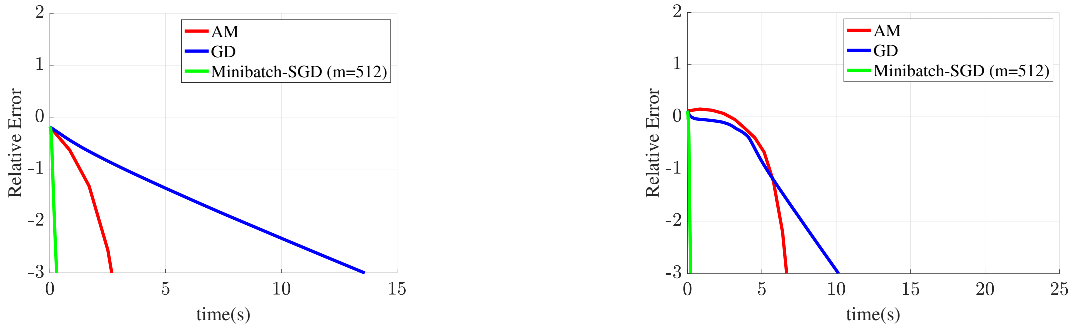
$$\min_{\pi \in \mathrm{Perm}([k])} \log_{10} \left(\sum_{j=1}^{k} \|\widehat{\boldsymbol{\beta}}_{\pi(j)} - \boldsymbol{\beta}_j^\star\|_2^2 / \sum_{j=1}^{k} \|\boldsymbol{\beta}_j^\star\|_2^2 \right)$$

with $\mathrm{Perm}([k])$ and $\{\widehat{\boldsymbol{\beta}}_j\}_{j=1}^k$ denoting the set of all possible permutations over $[k]$ and the estimated parameters, respectively. Our main result provides a theoretical analysis of SGD that explains this empirical observation.

1.1. Main results. We derive convergence analyses of GD and mini-batch SGD under the same covariate and noise assumptions in the previous work on AM by Ghosh et al. [12]. They assumed that covariates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are independent copies of a random vector $\boldsymbol{x}$ that satisfies the sub-Gaussianity and anti-concentration defined below.

Assumption 1.1 (Sub-Gaussianity). *The covariate distribution satisfies*

$$\|\langle \boldsymbol{v}, \boldsymbol{x} \rangle\|_{\psi_2} \leq \eta, \quad \forall \boldsymbol{v} \in \mathbb{S}^{d-1},$$



(a) Gaussian covariates (b) Uniform covariates

Figure 1: Convergence of estimators for noise-free max-affine regression ($k = 3$, $d = 500$, and $n = 8,000$).

where $\|\cdot\|_{\psi_2}$ and $\mathbb{S}^{d-1}$ denote the sub-Gaussian norm (i.e., see [44, Equation 2.13]) and the unit sphere in ℓ_2^d, respectively.

Assumption 1.2 (Anti-concentration). *The covariate distribution satisfies*

$$\sup_{w\in\mathbb{R},\boldsymbol{v}\in\mathbb{S}^{d-1}} \mathbb{P}((\langle \boldsymbol{v},\boldsymbol{x}\rangle + w)^2 \le \epsilon) \le (\gamma\epsilon)^{\zeta}, \quad \forall \epsilon > 0.$$

The class of covariate distributions by Assumptions 1.1 and 1.2 generalizes the standard independent and identically distributed Gaussian distribution. For example, the uniform and beta distributions satisfy Assumptions 1.1 and 1.2. Therefore, the theoretical result under this relaxed covariate model will apply to a wider range of applications. They also assumed that observations are corrupted with independent additive σ-sub-Gaussian noise.

This paper establishes the first theoretical analysis of GD and mini-batch SGD for max-affine regression. The following pseudo-theorem demonstrates that GD shows a local linear convergence under the above assumptions.

Theorem 1.3 (Informal). *Let $\boldsymbol{\beta}^\star \in \mathbb{R}^{k(d+1)}$ denote the column vector that collects all ground-truth parameters $(\boldsymbol{\theta}_j^\star, b_j^\star)_{j\in[k]}$. Given $\widetilde{O}(C_{\boldsymbol{\beta}^\star} kd(k^3 \vee \sigma^2))$ observations, a suitably initialized GD for max-affine regression converges linearly to an estimate of $\boldsymbol{\beta}^\star$ with ℓ_2-error scaling as $\widetilde{O}(\sigma k^2\sqrt{d/n})$, where $C_{\boldsymbol{\beta}^\star}$ is a constant that implicitly depends on k through $\boldsymbol{\beta}^\star$ but is independent of d.*

The error bound by this theorem improves upon the best-known result on max-affine regression achieved by AM [12, Theorem 2]. The error bound for AM is larger by a factor that grows at least as $k^{-1+2\zeta^{-1}}$. We also present an analogous analysis for SGD. A specification for the noise-free observation scenario is stated as follows.

Theorem 1.4 (Informal). *A suitably initialized mini-batch SGD for max-affine regression with $\widetilde{O}(C_{\boldsymbol{\beta}^\star} k^9 d)$ noise-free observations converges linearly to the ground truth $\boldsymbol{\beta}^\star$ for any batch size.*

The per-iteration cost of a mini-batch SGD with batch size m is $O(kmd)$, which is significantly lower than those for GD $O(knd)$ and of AM $O(knd^2)$. This implies the faster convergence of SGD in run time shown in Figure 1. We also observe that SGD empirically recovers the ground-truth parameters from fewer observations (see Figures 2 and 3).

1.2. Related Work. **Relation to phase retrieval and ReLU regression:** The max-affine model includes well-known models in signal processing and machine learning as special cases. The instance of (1.1) for $k = 2$ with $b_1^\star = b_2^\star = 0$ and $\boldsymbol{\theta}_1^\star = -\boldsymbol{\theta}_2^\star = \boldsymbol{\theta}^\star$ reduces to $y = |\langle \boldsymbol{x}, \boldsymbol{\theta}^\star \rangle|$, which corresponds to a measurement model in phase retrieval. Similarly, the rectified linear unit (ReLU) $y = \max(\langle \boldsymbol{x}, \boldsymbol{\theta}^\star \rangle, 0)$ is written in the form of (1.1) for $k = 2$ with $\boldsymbol{\theta}_1^\star = \mathbf{0}$ and $\boldsymbol{\theta}_2^\star = \boldsymbol{\theta}^\star$. A series of studies in [47, 38, 41, 40, 45, 25, 46, 43] has developed a statistical analysis of GD and SGD for phase retrieval and ReLU regression. It has been shown that for the noiseless case, GD and SGD converge linearly to a near-optimal estimate of the ground-truth parameters when the number of observations grows linearly with the ambient dimension d. In the context of bounded noise, GD converges to the ground truth within a radius determined by the noise level [47, 45]. However, it remained an open question whether GD is consistent under stochastic noise assumptions. Additionally, SGD in the presence of noise has not been thoroughly investigated yet. The main results of this paper address these questions on phase retrieval as a special case of max-affine regression.

Relation to convex regression: The max-affine model has also been adopted in parametric approaches to convex regression [30, 19, 18, 3, 1, 2, 36, 37, 35]. Let $f_\star : \mathbb{R}^d \to \mathbb{R}$ be an arbitrary convex function. The observations are given by $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ where $y_i = f_\star(\boldsymbol{x}_i)$ for all i in $[n]$. The nonparametric convex regression problem aims to estimate $f_\star$ by solving

$$\min_{f \in \mathcal{F}_{\mathrm{cvx}}} \sum_{i=1}^n (y_i - f(\boldsymbol{x}_i))^2, \tag{1.3}$$

where $\mathcal{F}_{\mathrm{cvx}}$ denotes the set of convex functions. Since f exists in the space of continuous real-valued functions on $\mathbb{R}^p$, the optimization problem in (1.3) is infinite-dimensional. A line of research [5, 3, 37] investigated the interpolation approach with a max-affine model in the form of

$$\widehat{f}(\boldsymbol{x}) = \max_{i \in [n]} \left(y_i + \boldsymbol{g}_i^\mathsf{T} (\boldsymbol{x} - \boldsymbol{x}_i) \right). \tag{1.4}$$

It provides a perfect interpolation of data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ with zero training error. For example, the interpolation is achieved by choosing $\boldsymbol{g}_i \in \partial f_\star(\boldsymbol{x}_i)$ for all $i \in [n]$. It has been show that the least squares estimator provides near-optimal generalization bounds relative to a matching minimax bound [28, 17, 1, 18, 27]. However, the minimax bound for the parametric model in (1.4) decays slowly due to the curse of dimensionality for a set of max affine with n segments. The least squares for the model in (1.4) is formulated as a quadratic program (QP) [5, Section 6.5.5]. However, off-the-shelf interior-point methods do not scale to large instances of this QP due to the high computational cost $O(d^4 n^5)$ [30, 19].

The k-max-affine model in (1.1) is considered as an alternative compact parametrization to approximate convex regression. The worst-case error in approximating d-variate Lipschtiz convex functions on a bounded domain by a k-max-affine model decays as $O(k^{-2/d})$

[1, Lemma 5.2]. However, data in practical applications such as aircraft wing design, wage prediction, and pricing stock options are often well approximated by the k-max-affine model with small k (e.g., [19, Section 6], [1, Section 7]). Unlike the interpolation approach to convex regression, if the compact model fits data in applications, the estimation error decays much faster in n.

Max-linear regression in the presence of deterministic noise: A special instance of (1.1) with $b_j^\star = 0$ for $j \in [k]$ is called the max-linear model. A convex optimization method to max-linear regression obtained with an initial estimate has been studied under the standard Gaussian covariate assumption and deterministic noise [26]. They empirically showed that the convex estimator outperforms the existing methods in the presence of outliers.

1.3. Organizations and Notations.

The rest of the paper is organized as follows: Section 2 formulates the least squares estimator for max-affine regression, describes the GD algorithm and presents the convergence analysis of GD. Section 3 describes a mini-batch SGD for max-affine regression and provides its convergence analysis. Section 4 presents numerical results to compare the empirical performance of GD, SGD, and AM for max-affine regression. Finally, Section 5 summarizes the contributions and discusses future directions.

Boldface lowercase letters denote column vectors, and boldface capital letters denote matrices. The concatenation of two column vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ is denoted by $[\boldsymbol{a};\ \boldsymbol{b}]$. The subvector of $\boldsymbol{a} \in \mathbb{R}^{d+1}$ with the first d entries will be denoted by $(\boldsymbol{a})_{1:d}$. Various norms are used throughout the paper. We use $\|\cdot\|$, $\|\cdot\|_{\mathrm{F}}$, $\|\cdot\|_2$, and $\|\cdot\|_{\psi_2}$ to denote the spectral norm, Frobenius norm, Euclidean norm, and sub-Gaussian norm respectively. Moreover, B_2^d and $\mathbb{S}^{d-1}$ will denote the d-dimensional unit ball and unit sphere with respect to the Euclidean norm. For two scalars q and d, we write $q \lesssim p$ if there exists an absolute constant $C > 0$ such that $q \leq Cp$. We use $C, C_1, C_2, \dots$ and $c, c_1, c_2, \dots$ to denote absolute constants that may vary from line to line. We adopt the big-O notation so that $q \lesssim p$ is alternatively written as $q = O(p)$. With a tilde on top of O, we ignore logarithmic factors. For brevity, the shorthand notation $[n]$ denotes the set $\{1, \dots, n\}$ for $n \in \mathbb{N}$. Moreover, $a \vee b$ and $a \wedge b$ will denote $\max(a, b)$ and $\min(a, b)$ for $a, b \in \mathbb{R}$.

2. Convergence analysis of gradient descent.

We first formulate the least squares estimator for max-affine regression and derive the gradient descent algorithm. For brevity, let $\boldsymbol{\xi} := [\boldsymbol{x};\ 1] \in \mathbb{R}^{d+1}$ and $\boldsymbol{\beta}_j := [\boldsymbol{\theta}_j;\ b_j] \in \mathbb{R}^{d+1}$. Then the model in (1.1) is rewritten as

$$y = \max_{j \in [k]} \langle \boldsymbol{\xi}, \boldsymbol{\beta}_j^\star \rangle + \text{noise}. \tag{2.1}$$

The least squares estimator minimizes the quadratic loss function given by

$$\ell(\boldsymbol{\beta}) := \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \max_{j \in [k]} \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j \rangle \right)^2, \tag{2.2}$$

where $\boldsymbol{\beta} = [\boldsymbol{\beta}_1;\ \dots;\ \boldsymbol{\beta}_k] \in \mathbb{R}^{k(d+1)}$.

The gradient descent algorithm iteratively updates the estimate by

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t - \mu \nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}^t),$$

where $\mu > 0$ denotes a step size. A generalized gradient [21] of the cost function in (2.2) with respect to the jth block $\boldsymbol{\beta}_j$ is written as

$$\nabla_{\boldsymbol{\beta}_j} \ell(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j\}} \left(\max_{j \in [k]} \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j \rangle - y_i \right) \boldsymbol{\xi}_i, \tag{2.3}$$

where $\mathcal{C}_1, \ldots, \mathcal{C}_k$ are defined by $\boldsymbol{\beta}$ as

$$\mathcal{C}_j := \{ \boldsymbol{w} \in \mathbb{R}^d \ : \ \langle [\boldsymbol{w};\ 1], \boldsymbol{\beta}_j - \boldsymbol{\beta}_l \rangle > 0, \ \forall l < j, \ \langle [\boldsymbol{w};\ 1], \boldsymbol{\beta}_j - \boldsymbol{\beta}_l \rangle \geq 0, \ \forall l > j \}. \tag{2.4}$$

The set $\mathcal{C}_j$ contains all inputs maximizing the jth linear model.[1] Note that each $\mathcal{C}_j$ is determined by $k-1$ half spaces given by the pairwise difference of the jth linear model and the others.

We show that the expression in (2.3) provides a valid generalized gradient of $\ell(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}_\ell$. We apply the chain rule on the generalized gradient [21]. The cost function in (2.2) is the composition $\varrho \circ F$ where

$$\varrho((z_i)_{i=1}^n) = \frac{1}{2n} \sum_{i=1}^{n} z_i^2$$

and $\boldsymbol{\beta} \mapsto F(\boldsymbol{\beta}) = (f_i(\boldsymbol{\beta}))_{i=1}^n$ with

$$f_i(\boldsymbol{\beta}) = \left| \max_{j \in [k]} \langle \boldsymbol{\beta}_j, \boldsymbol{\xi}_i \rangle - y_i \right|, \quad i \in [n].$$

Since each max-affine function f_i is regular at each point of the domain, the equality in [21, Eq. (5.7)] holds and it characterizes the generalized gradient of ℓ as

$$\nabla_{\boldsymbol{\beta}_\ell} \ell(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \left(\max_{j \in [k]} \langle \boldsymbol{\beta}_j, \boldsymbol{\xi}_i \rangle - y_i \right) \cdot \nabla_{\boldsymbol{\beta}_\ell} \left(\max_{j \in [k]} \langle \boldsymbol{\beta}_j, \boldsymbol{\xi}_i \rangle \right).$$

Since a sub-gradient of a convex function is a generalized gradient [6], it suffices to show that $\mathbb{1}_{\{\boldsymbol{x}_i \in C_\ell\}} \boldsymbol{\xi}_i$ is a sub-gradient of the convex function $\nabla_{\boldsymbol{\beta}_\ell} \left(\max_{j \in [k]} \langle \boldsymbol{\beta}_j, \boldsymbol{\xi}_i \rangle \right)$. To this end, we verify that the following inequality holds for all $i \in [n]$:

$$\max \left(\langle \boldsymbol{\beta}_\ell + \boldsymbol{h}, \boldsymbol{\xi}_i \rangle, \max_{j \neq \ell \in [k]} \langle \boldsymbol{\beta}_j, \boldsymbol{\xi}_i \rangle \right) - \max_{j \in [k]} \langle \boldsymbol{\beta}_j, \boldsymbol{\xi}_i \rangle \geq \mathbb{1}_{\{\boldsymbol{x}_i \in C_\ell\}} \langle \boldsymbol{h}, \boldsymbol{\xi}_i \rangle, \quad \forall \boldsymbol{h} \in \mathbb{R}^{d+1}. \tag{2.5}$$

Let $i \in [n]$ be arbitrarily fixed. First, we consider the case when ℓ is a maximizer in the max-affine function in (2.1) at $\boldsymbol{\xi}_i$. Then we have $\langle \boldsymbol{\beta}_\ell, \boldsymbol{\xi}_i \rangle = \max_{j \in [k]} \langle \boldsymbol{\beta}_j, \boldsymbol{\xi}_i \rangle$ and $\mathbb{1}_{\{\boldsymbol{x}_i \in C_\ell\}} = 1$. Therefore, (2.5) holds since

$$\max \left(\langle \boldsymbol{\beta}_\ell + \boldsymbol{h}, \boldsymbol{\xi}_i \rangle, \max_{j \neq \ell \in [k]} \langle \boldsymbol{\beta}_j, \boldsymbol{\xi}_i \rangle \right) \geq \langle \boldsymbol{\beta}_\ell + \boldsymbol{h}, \boldsymbol{\xi}_i \rangle, \quad \forall \boldsymbol{h} \in \mathbb{R}^{d+1}.$$

[1] In case of a tie when multiple linear models attain the maximum for a given sample, we assign the sample to the smallest maximizing index. Since the event of duplicate maximizing indices will happen with probability 0 for any absolutely continuous probability measure on $\boldsymbol{x}_i$s, the choice of a tie-break rule does not affect the analysis.

Next, we assume that ℓ is not a maximizer. Then $\mathbb{1}_{\{\boldsymbol{x}_i \in C_\ell\}} = 0$ and there exists $\ell' \in [k] \setminus \{\ell\}$ such that $\langle \boldsymbol{\beta}_{\ell'}, \boldsymbol{\xi}_i \rangle = \max_{j\in[k]} \langle \boldsymbol{\beta}_j, \boldsymbol{\xi}_i \rangle > \langle \boldsymbol{\beta}_\ell, \boldsymbol{\xi}_i \rangle$. Therefore, (2.5) is also satisfied since

$$\max\left(\langle \boldsymbol{\beta}_\ell + \boldsymbol{h}, \boldsymbol{\xi}_i \rangle, \langle \boldsymbol{\beta}_{\ell'}, \boldsymbol{\xi}_i \rangle\right) \geq \langle \boldsymbol{\beta}_{\ell'}, \boldsymbol{\xi}_i \rangle, \quad \forall \boldsymbol{h} \in \mathbb{R}^{d+1}.$$

Then the generalized gradient $\nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta})$ is obtained by concatenating $\{\nabla_{\boldsymbol{\beta}_j} \ell(\boldsymbol{\beta})\}_{j=1}^k$ by

$$\nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) = \sum_{j=1}^k \boldsymbol{e}_j \otimes \nabla_{\boldsymbol{\beta}_j} \ell(\boldsymbol{\beta}),$$

where $\boldsymbol{e}_j \in \mathbb{R}^k$ denotes the jth column of the k-by-k identity matrix $\boldsymbol{I}_k$ for $j \in [k]$. Moreover, $\ell(\boldsymbol{\beta})$ is differentiable except on a set of measure zero, with a slight abuse of terminology, $\nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta})$ is referred to as the "gradient".

Next, we present a convergence analysis of the gradient descent estimator. The analysis depends on a set of geometric parameters of the ground-truth model. The first parameter $\pi_{\min}$ describes the minimum portion of observations corresponding to the linear model which achieved the maximum least frequently. It is formally defined as a lower bound on the probability measure on the smallest partition set, i.e.

$$\min_{j\in[k]} \mathbb{P}(\boldsymbol{x} \in \mathcal{C}_j^\star) \geq \pi_{\min}, \tag{2.6}$$

where $\mathcal{C}_1^\star, \ldots, \mathcal{C}_k^\star$ are polytopes determined by

$$\mathcal{C}_j^\star := \{\boldsymbol{w} \in \mathbb{R}^d \ : \ \langle [\boldsymbol{w};\ 1], \boldsymbol{\beta}_j^\star - \boldsymbol{\beta}_l^\star \rangle > 0,\ \forall l < j,\ \langle [\boldsymbol{w};\ 1], \boldsymbol{\beta}_j^\star - \boldsymbol{\beta}_l^\star \rangle \geq 0,\ \forall l > j\}. \tag{2.7}$$

The next parameter κ quantifies the separation between all pairs of distinct linear models in (1.1) so that the pairwise distance on two distinct linear models satisfy

$$\min_{j' \neq j} \|(\boldsymbol{\beta}_j^\star)_{1:d} - (\boldsymbol{\beta}_{j'}^\star)_{1:d}\|_2 \geq \kappa. \tag{2.8}$$

Next, we present a convergence analysis of the gradient descent estimator. The analysis depends on a set of geometric parameters of the ground-truth model. The first parameter $\pi_{\min}$ describes the minimum portion of observations corresponding to the linear model which achieved the maximum least frequently. It is formally defined as a lower bound on the probability measure on the smallest partition set, i.e.

$$\min_{j\in[k]} \mathbb{P}(\boldsymbol{x} \in \mathcal{C}_j^\star) \geq \pi_{\min}, \tag{2.9}$$

where $\mathcal{C}_1^\star, \ldots, \mathcal{C}_k^\star$ are polytopes determined by

$$\mathcal{C}_j^\star := \{\boldsymbol{w} \in \mathbb{R}^d \ : \ \langle [\boldsymbol{w};\ 1], \boldsymbol{\beta}_j^\star - \boldsymbol{\beta}_l^\star \rangle > 0,\ \forall l < j,\ \langle [\boldsymbol{w};\ 1], \boldsymbol{\beta}_j^\star - \boldsymbol{\beta}_l^\star \rangle \geq 0,\ \forall l > j\}. \tag{2.10}$$

The next parameter κ quantifies the separation between all pairs of distinct linear models in (1.1) so that the pairwise distance on two distinct linear models satisfy

$$\min_{j' \neq j} \|(\boldsymbol{\beta}_j^\star)_{1:d} - (\boldsymbol{\beta}_{j'}^\star)_{1:d}\|_2 \geq \kappa. \tag{2.11}$$

Our main result in the following theorem presents a local linear convergence of the gradient descent estimator uniformly over all $\boldsymbol{\beta}^\star$ satisfying (2.10) and (2.11).

Theorem 2.1. *Let $\delta \in (0, 1/e)$, $y_i = \max_{j\in[k]}\langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j^\star\rangle + z_i$ for $i \in [n]$ with $\boldsymbol{\xi}_i = [\boldsymbol{x}_i;\ 1]$, and $\{z_i\}_{i=1}^n$ being additive σ-sub-Gaussian noise independent from everything else. Suppose that Assumptions 1.1 and 1.2 hold.*[2] *Then there exist absolute constants $C, C', R > 0$, and $\nu \in (0, 1)$, for which the following statement holds with probability at least $1-\delta$: If the initial estimate $\boldsymbol{\beta}^0$ belongs to a neighborhood of $\boldsymbol{\beta}^\star$ given by*

$$\mathcal{N}(\boldsymbol{\beta}^\star) := \left\{ \boldsymbol{\beta} \in \mathbb{R}^{k(d+1)} \ :\ \max_{j\in[k]} \|\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^\star\|_2 \le \kappa\rho \right\} \tag{2.12}$$

with

$$\rho := \frac{R\pi_{\min}^{\zeta^{-1}(1+\zeta^{-1})}}{4k^{\zeta^{-1}}} \cdot \log^{-1/2}\left(\frac{k^{\zeta^{-1}}}{R\pi_{\min}^{\zeta^{-1}(1+\zeta^{-1})}} \right) \wedge \frac{1}{4}, \tag{2.13}$$

then for all $\boldsymbol{\beta}^\star$ satisfying (2.9) and (2.11), the sequence $\left(\boldsymbol{\beta}^t\right)_{t\in\mathbb{N}}$ by the gradient descent method with a constant step size satisfies

$$\left\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^\star\right\|_2 \le \nu^t \left\|\boldsymbol{\beta}^0 - \boldsymbol{\beta}^\star\right\|_2 + C'\sigma k \frac{\sqrt{k\left(kd\log(n/d) + \log(k/\delta)\right)}}{\sqrt{n}}, \quad \forall t \in \mathbb{N}, \tag{2.14}$$

provided that

$$n \ge C\pi_{\min}^{-2(1+\zeta^{-1})} \cdot \left(k^{1.5}\pi_{\min}^{-(1+\zeta^{-1})} \vee \frac{\sigma}{\kappa\rho} \right)^2 \cdot \left(kd\log(n/d) + \log(k/\delta)\right). \tag{2.15}$$

Proof. See Section SM3. ■

Theorem 2.1 demonstrates that the GD estimator with a constant step size converges linearly to a neighborhood of the ground-truth parameter of radius $\widetilde{O}\left(\sigma^2 k^4 d/n\right)$. The number of sufficient observations to invoke this convergence result scales linearly in d and is proportional to a polynomial in $\pi_{\min}^{-1}$ and k. This result implies the consistency of the gradient descent estimator. To compare Theorem 2.1 to the analogous result for AM under the same covariate and noise models [13, Theorem 1], we have the following remarks in order.

- First, the final estimation error by (2.14) with $t \to \infty$ is smaller than that by [13, Theorem 1] by being independent of $\pi_{\min}^{-1}$, which grows at least proportional to k. A larger estimation error bound in their result is due to the analysis of the least squares update, wherein the smallest singular value of the design matrix of each linear model is utilized. These quantities do not appear in the analysis of the gradient descent update.
- Second, the convergence parameter ν in (2.14) is smaller than 3/4 for AM[3], which might result in a slower convergence of GD in iteration count. The convergence speed

[2] To simplify the presentation, we assume that the parameters η, ζ, γ in Assumptions 1.1 and 1.2 are fixed numerical constants in the statement and proof of Theorem 2.1. Therefore, any constant determined only by η, ζ, γ will be treated as a numerical constant.

[3] As shown in the proof in Section SM3, the parameter ν is given as $\nu = (1-\mu\lambda)$ by (SM3.19). The quantity $\mu\lambda$ is determined by (SM3.8) and (SM3.29) as a function of $\pi_{\min}$, $\pi_{\max}$, and ζ so that it decreases in k and $\pi_{\min}^{-1}$.

issue becomes significant for large k and $\pi_{\min}^{-1}$. For example, in the illustration by Figure 1, GD shows a slower convergence in run time despite the lower per-iteration cost $O(knd)$, which is lower than that of AM $O(knd^2)$ by a factor of d. However, as discussed in Section 3, the slow convergence of GD can be improved by modifying the algorithm into a (mini-batch) SGD.

- Third, the sample complexity results by Theorem 2.1 and [13, Theorem 1] are qualitatively comparable. There were mistakes in the proof of [13, Theorem 1]. We think that their result could be corrected with an increased order of dependence in their sample complexity on k and $\pi_{\min}$ (see Section SM5 for a detailed discussion).
- Lastly, regarding the proof technique, we adapt and improve the strategy by Ghosh et al. [12, 13]. Note that the subgradient of the loss function in (2.3) involves clustering of covariates with respect to maximizing linear models such as (2.4), which also arises in alternating minimization. Due to this similarity, key quantities in the analysis have been estimated in [12, 13]. We provide sharpened estimates via different techniques. For example, Lemma SM2.3 provides a tighter bound than [12, Lemma 7] by a factor of $\alpha^{\zeta^{-1}}$ for a scalar $\alpha \in (0,1)$.

Theorem 2.1 also provides an auxiliary result. As a direct consequence of Theorem 2.1, we obtain an upper bound on the prediction error, which is defined by

$$\mathcal{E}(\widehat{\boldsymbol{\beta}}) := \mathbb{E}\left(\max_{j\in[k]}\langle \boldsymbol{\xi}, \widehat{\boldsymbol{\beta}}_j\rangle - \max_{j\in[k]}\langle \boldsymbol{\xi}, \boldsymbol{\beta}_j^\star\rangle\right)^2,$$

where $\widehat{\boldsymbol{\beta}} = [\widehat{\boldsymbol{\beta}}_1; \ldots; \widehat{\boldsymbol{\beta}}_k]$ denotes the estimated parameter vector by GD. Since the quadratic cost function in (1.2) is 1-Lipschitz with respect to the ℓ_2 norm, it follows that the prediction error $\mathcal{E}(\widehat{\boldsymbol{\beta}})$ is also bounded by $\widetilde{O}(\sigma^2 k^3 d/n)$ as in (2.14) with $t \to \infty$.

A limitation of Theorem 2.1 is that its local convergence analysis requires an initialization within a specific neighborhood of the ground-truth parameter. To obtain the desired initial estimate, one may use spectral initialization by [14, Algorithm 2, 3], which consists of dimensionality reduction followed by a grid search. They provided a performance guarantee of a spectral initialization scheme under the standard Gaussian covariate assumption [14, Theorems 2 and 3]. Therefore, the reduction of Theorem 2.1 to the Gaussian covariate case combined with [14, Theorems 2 and 3] provides a global convergence analysis of GD, which is comparable to that for alternating minimization [14]. Even in this case, the number of sufficient samples for the success of spectral initialization overwhelms that for the subsequent gradient descent step. Since multiple steps of their analysis critically depend on the Gaussianity, it remains an open question whether the result on the spectral initialization generalizes to the setting by Assumptions 1.1 and 1.2.

3. Convergence analysis of mini-batch SGD.

SGD is an optimization method that updates parameters using a single or a small batch of randomly selected data point(s) instead of the entire dataset. SGD converges faster in run time than GD due to its significantly lower per-iteration cost. In particular, when applied to max-affine regression, SGD empirically outperforms GD and AM in both sample complexity and convergence speed (see Figures 1 to 3). In this section, we present an accompanying theoretical convergence analysis of mini-batch SGD for max-affine regression. The update rule of a mini-batch SGD with batch size m for max-affine regression is described as follows. For each iteration index $t \in \mathbb{N}$, let I_t be a multiset of m randomly selected indices with replacement so that the entries of I_t are independent copies of a uniform random variable in $[n]$. A mini-batch SGD iteratively updates the estimate by

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t - \mu \frac{1}{m} \sum_{i \in I_t} \nabla_{\boldsymbol{\beta}} \ell_i(\boldsymbol{\beta}^t),$$

where

$$\ell_i(\boldsymbol{\beta}) := \frac{1}{2} \left(y_i - \max_{j \in [k]} \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j \rangle \right)^2, \quad i \in [n].$$

Then the following theorem presents a local linear convergence of SGD.

Theorem 3.1. *Under the hypothesis of Theorem 2.1, there exist absolute constants $C, C' > 0$ and $c, \nu \in (0,1)$, for which the following statement holds with probability at least $1-\delta$: For all $\boldsymbol{\beta}^\star$ satisfying (2.10) and (2.11), if the initial estimate $\boldsymbol{\beta}^0$ belongs to $\mathcal{N}(\boldsymbol{\beta}^\star)$ defined in (2.12), n satisfies (2.15), and m satisfies*

$$m \geq C \cdot \left(\frac{\sigma}{\kappa \rho} \right)^2 \cdot \left(d + \log(k/\delta) \right), \tag{3.1}$$

then the sequence $\left(\boldsymbol{\beta}^t\right)_{t \in \mathbb{N}}$ by the mini-batch SGD with batch size m and step size $\mu = c\,(1 \wedge m/(d + \log(n/\delta)))$ satisfies

$$\begin{aligned} \mathbb{E}_{I_t} \left\| \boldsymbol{\beta}^t - \boldsymbol{\beta}^\star \right\|_2 &\leq \left(1 - \left(1 \wedge \frac{m}{d + \log(n/\delta)} \right) c\nu \right)^t \left\| \boldsymbol{\beta}^0 - \boldsymbol{\beta}^\star \right\|_2 \\ &\quad + C' \sigma k \sqrt{\left(\frac{d + \log(n/\delta)}{m} \vee \frac{kd \log(n/d) + \log(1/\delta)}{n} \right)}, \quad \forall t \in \mathbb{N}. \end{aligned} \tag{3.2}$$

Proof. See Section SM4. ∎

Theorem 3.1 establishes linear convergence of mini-batch SGD in expectation to the ground-truth parameters within error $\widetilde{O}(\sigma^2 k^2 \,(d/m \vee kd/n))$. The local linear convergence applies uniformly over all $\boldsymbol{\beta}^\star$ satisfying (2.10) and (2.11). In general, the convergence rate of SGD is much slower even with strong convexity [33, 4, 20]. However, in a special case where the cost function is in the form of $\sum_{i=1}^n \ell_i(\boldsymbol{\beta})$, smooth, and strongly convex, if $\boldsymbol{\beta}^\star$ is the minimizer of all summands $\{\ell_i(\boldsymbol{\beta})\}_{i=1}^n$, then SGD converges linearly to $\boldsymbol{\beta}^\star$ [32, Theorem 2.1]. The convergence analysis in Theorem 3.1 can be considered along with this result. The cost function in (2.2) in the noiseless case satisfies the desired properties locally near the ground truth, whence establishes the local linear convergence of SGD.

Theorem 3.1 also explains how the batch size m affects the final estimation error by (3.2) with $t \to \infty$. Let n and m satisfy (2.15) and (3.1) so that Theorem 3.1 is invoked. Under this condition, one can still choose m and n so that $m \lesssim n/k$. Then the $\widetilde{O}(\sigma^2 k^2 d/m)$ term determined by the batch size m dominates the final estimation error. In this regime, the SGD estimator is not consistent since the estimation error $\widetilde{O}(\sigma^2 k^2 d/m)$ does not vanish with increasing n. This result implies the trade-off between the convergence speed and the final estimation error determined by the batch size.

Furthermore, since the condition on m in (3.1) becomes trivial when $\sigma = 0$, we obtain a stronger result in the noiseless case given by the following corollary.

Corollary 3.2. *Let $\delta, \delta' \in (0,1)$, and $\epsilon > 0$ fixed. Suppose that the hypothesis of Theorem 3.1 holds. If $t \geq (\log(1/\epsilon) + \log(1/\delta)) \left(1 \vee \frac{d+\log(n/\delta)}{m}\right) 1/\nu$, then*

$$\left\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^\star\right\|_2 \leq \epsilon \|\boldsymbol{\beta}^0 - \boldsymbol{\beta}^\star\|_2$$

holds with probability at least $1 - \delta - \delta'$.

Proof. By Theorem 3.1, (3.2) holds with probability at least $1-\delta$. By applying Markov's inequality, we have

$$\mathbb{P}\left(\left\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^\star\right\|_2 \geq \epsilon\|\boldsymbol{\beta}^0 - \boldsymbol{\beta}^\star\|_2\right) \leq \frac{\mathbb{E}_{I_t}\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^\star\|_2}{\epsilon\|\boldsymbol{\beta}^0 - \boldsymbol{\beta}^\star\|_2} \leq \frac{\left(1 - \left(1 \wedge \frac{m}{d+\log(n/\delta)}\right)\nu\right)^t}{\epsilon} \leq \delta',$$

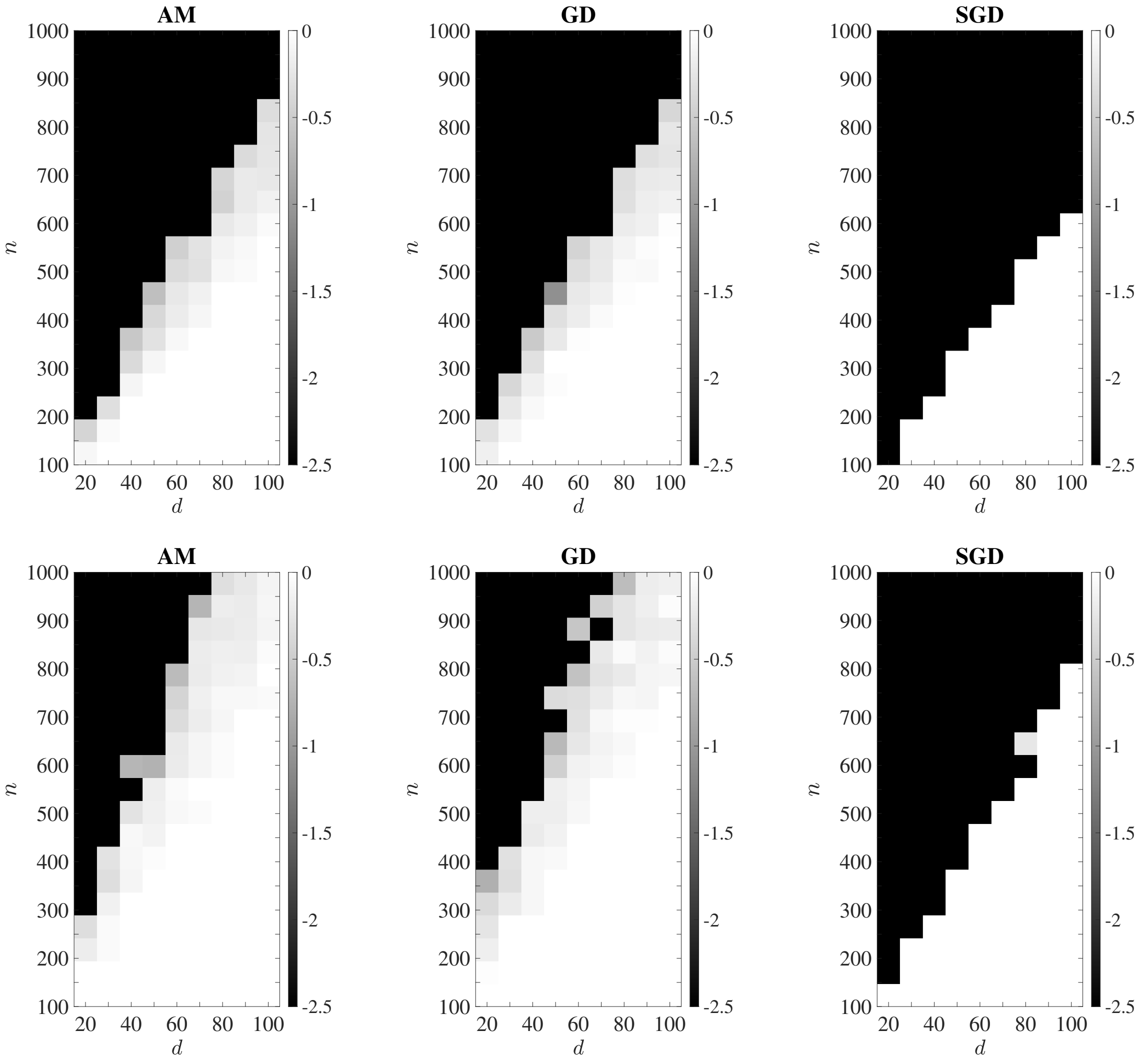
where the second and third inequalities hold by (3.2) and assumption on t respectively. ∎

Corollary 3.2 presents the convergence of SGD with high probability, which is stronger than the convergence in expectation. Furthermore, there is no requirement on the batch size in invoking Corollary 3.2. This result is analogous to the recent theoretical analysis of phase retrieval by randomized Kaczmarz [41] and SGD [40].

4. Numerical results.

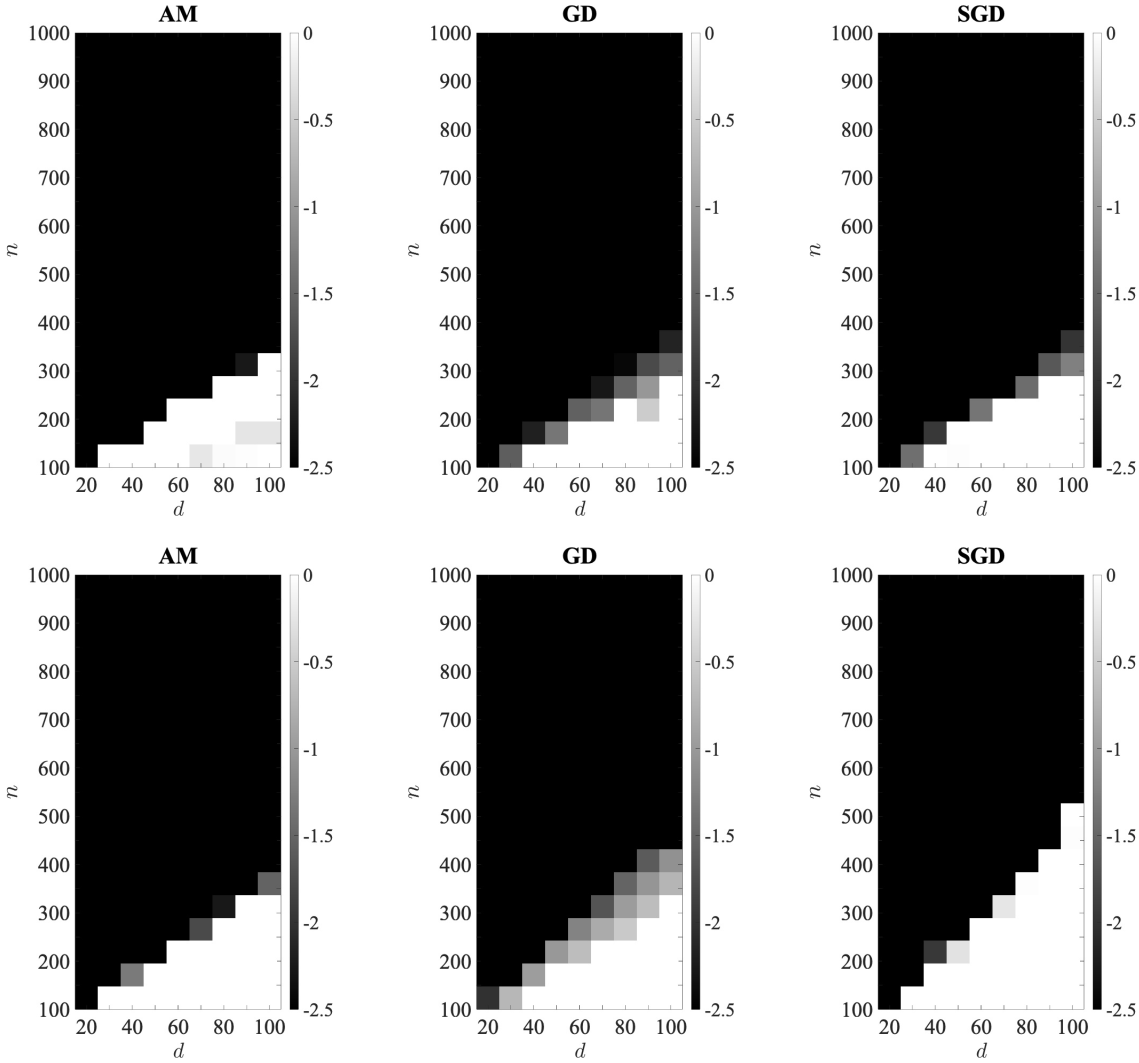
We study the empirical performance of GD and mini-batch SGD for max-affine regression. The performance of these first-order methods is compared to AM [14]. All of these algorithms start from the spectral initialization by Ghosh et al. [14]. We use a constant step size 0.5 for GD. The step size for SGD is set to $\frac{1\wedge(m/d)}{2}$ adaptive to the batch size. According to our covariate assumptions in Assumption 1.1 and Assumption 1.2, we consider the following two scenarios; The first scenario involves Gaussian covariates, where $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are generated as independent samples from a random vector following Normal$(\mathbf{0}, \mathbf{I}_d)$. The other scenario involves a uniform distribution, where $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are generated as independent samples from a random vector following Unif$[-\sqrt{3}, \sqrt{3}]^{\otimes d}$, which is also considered in the numerical setting in [12]. We use spectral initialization for the Gaussian covariate model [12], while for the uniform distribution case, we apply the multiple-restart random initialization method [1].

First, we observe the performance of the three estimators for the exact parameter recovery in the noiseless case. In this experiment, the ground-truth parameters $\boldsymbol{\theta}_1^\star, \ldots, \boldsymbol{\theta}_k^\star$ are generated as k random pairwise orthogonal vectors with $k < d$, and the offset terms are set to 0, i.e., $b_j^\star = 0$ for all $j \in [k]$. By the construction, the probability assigned to the maximizer set of each linear model will be approximately $\frac{1}{k}$. In other words, the parameters $\pi_{\max}$ and $\pi_{\min}$ of the ground truth concentrate around $\frac{1}{k}$ where $\pi_{\min}$ is defined in (2.9) and $\pi_{\max} := \max_{j\in[k]} \mathbb{P}(\boldsymbol{x} \in \mathcal{C}_j^\star)$. Furthermore, due to the orthogonality, the pairwise distance satisfies $\|\boldsymbol{\theta}_j^\star - \boldsymbol{\theta}_{j'}^\star\|_2 = \sqrt{2}$ for all $j \neq j' \in [k]$. Consequently, the sample complexity results for GD and SGD by Theorem 2.1 and Theorem 3.1 simplify to an easy-to-interpret expression $\widetilde{O}(k^{16}d)$ that involves only k and d for both Gaussian and uniform distribution scenarios. The sample complexity result on AM [12] simplifies similarly.



(a) Gaussian covariate

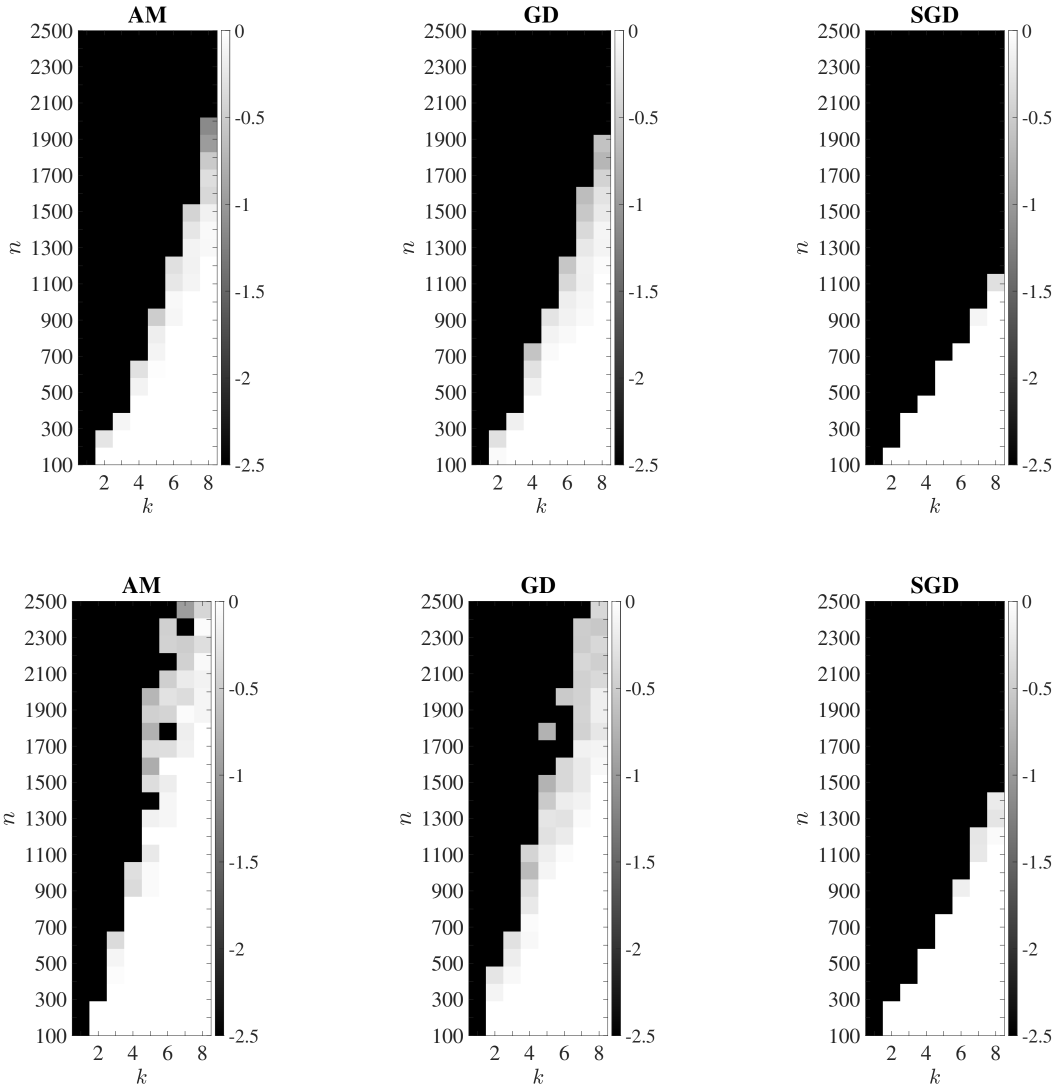
Figures 2a and 3a illustrate the empirical phase transition by the three estimators through Monte Carlo simulations under the Gaussian covariate model. The median and the 90th percentile of 50 random trials are displayed. In these figures, the transition occurs when the sample size n becomes larger than a threshold that depends on the ambient dimension d and the number of linear models k. Figure 2a shows that the threshold for both estimators increases linearly with d for fixed k. This observation is consistent with the sample complexity by Theorem 2.1 and Theorem 3.1. A complementary view is presented in Figure 3a for varying k and fixed d. The thresholds in Figure 3a for GD and SGD are almost linear in k when d is fixed to 50, which scales slower than the corresponding sample complexity results in Theorem 2.1 and Theorem 3.1. A similar discrepancy between theoretical and empirical phase transitions has been observed for AM [12, Appendix L]. We also observe that mini-batch SGD



(b) Uniform covariate

Figure 2: Phase transition of estimation error per the number of observations n and the ambient dimension d in the noiseless case (The number of linear models k and the batch size m are set to 3 and 64, respectively). The first row and the second row respectively show the median and the 90th percentile of estimation errors in 50 trials.

outperforms GD and AM with a lower threshold for phase transition. It has been shown that the inherent random noise in the gradient helps the estimator to escape saddle points or local minima [23, 8]. This explains why SGD recovers the parameters with fewer samples than GD. We also note that the relative performance among the three estimators remains similar in both the median and the 90th percentile. This shows that SGD for noiseless max-affine

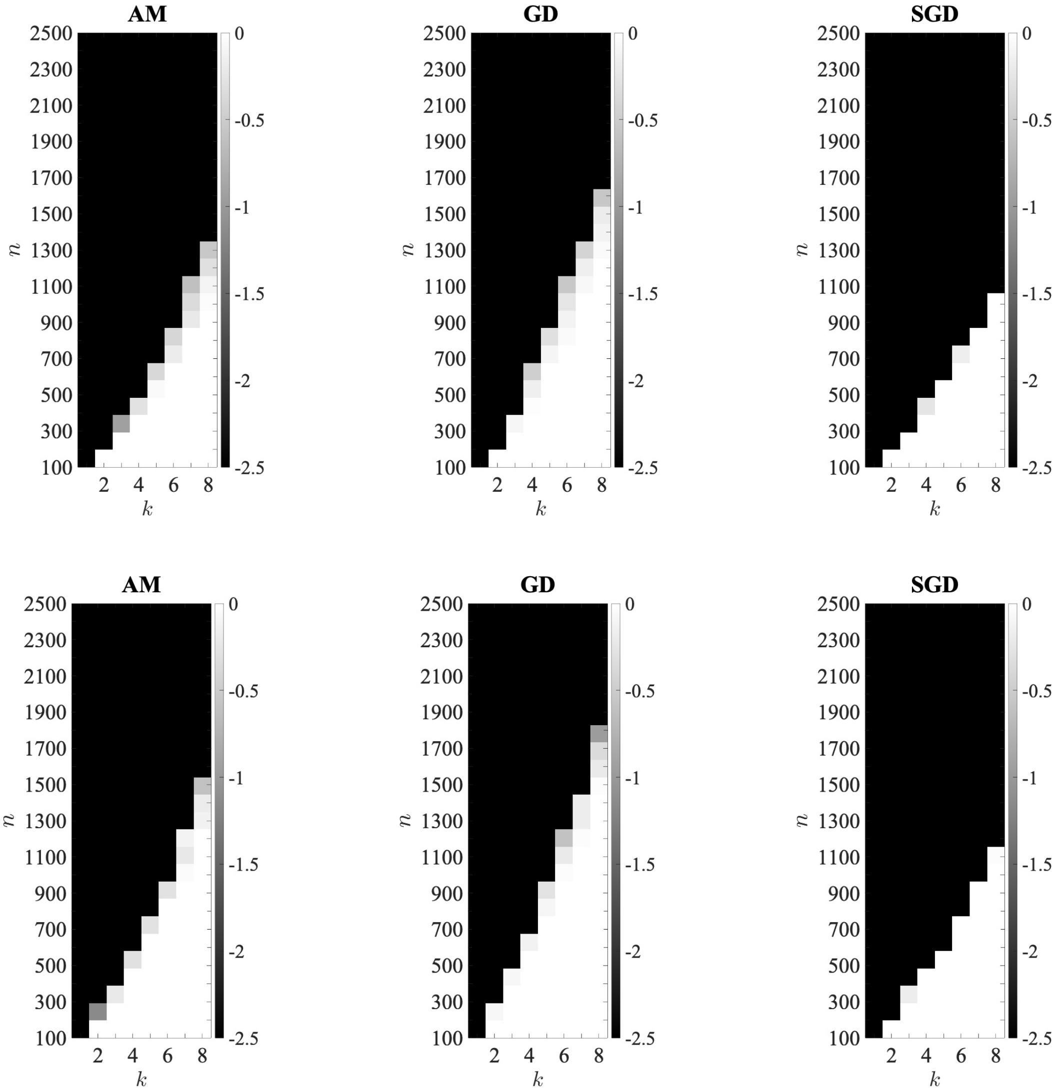


(a) Gaussian covariate

regression does not suffer from a large variance, which corroborates the result in Corollary 3.2.

The phase transition boundaries in Figures 2b and 3b are higher with a larger success regime relative to the corresponding results in Figures 2a and 3a. Recall that GD/SGD with the multiple-restart random initialization involves multiple runs of GD/SGD. The performance improvement is obtained at the cost of higher computational cost proportional to the number of repetitions.

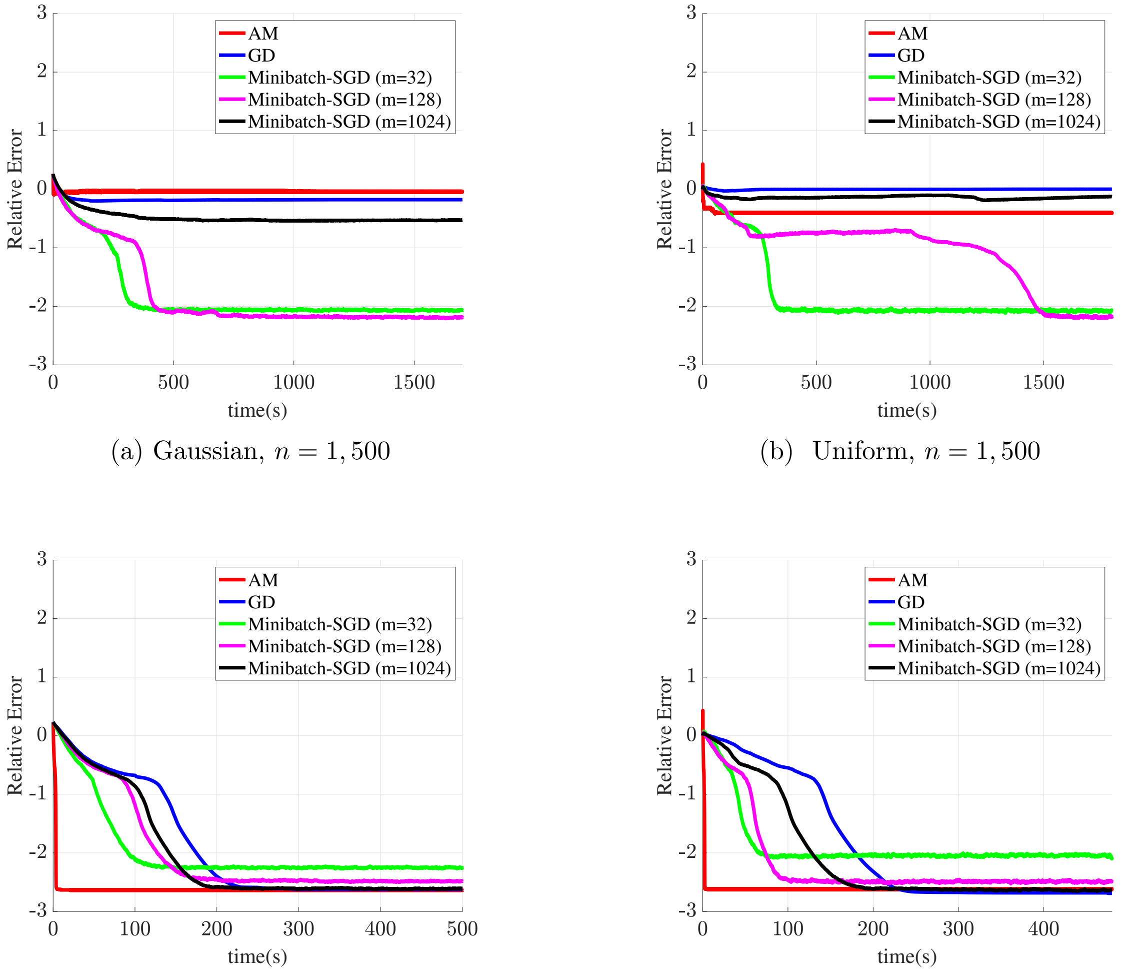
Figures 4 and 5 study the estimation error by mini-batch SGD under zero-mean Gaussian noise with standard deviation $\sigma = 0.1$ in three different scenarios. In Figure 4, we focus on observing how the batch size m affects the convergence speed and the estimation error. Figure 4a and Figure 4b consider the scenario where the spectral method provides a poor



(b) Uniform covariate

Figure 3: Phase transition of estimation error per number of observations n and number of linear models k in the noiseless case (The ambient dimension d and mini-batch size m are set to 50 and 64 respectively). The first row and the second row respectively show the median and the 90th percentile of estimation errors in 50 trials.

initialization due to a small number of observations. Consequently, GD and AM fail to provide a low estimation error. In contrast, mini-batch SGD with a small batch size ($m = 32$ or $m = 128$) relative to the total number of samples ($n = 1,500$) converges to a small estimation error ($< 10^{-2}$). In other words, there exists a trade-off between the convergence speed and the estimation error determined by the batch size m. SGD with $m = 128$ converges



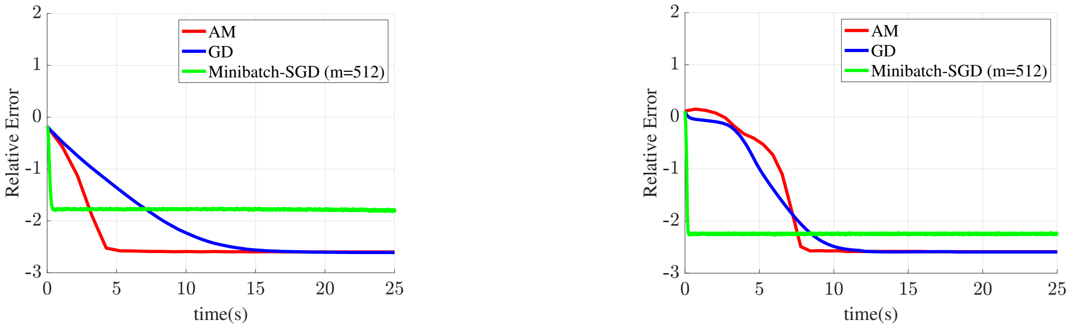
(a) Gaussian, $n = 1,500$ (b) Uniform, $n = 1,500$

(c) Gaussian, $n = 3,000$ (d) Uniform, $n = 3,000$

Figure 4: Convergence of estimators for max-affine regression under additive white Gaussian noise of variance $\sigma^2 = 0.01$ ($k = 8$ and $d = 50$). Comparison between Gaussian and Uniform covariates.

slower to a smaller error than SGD with $m = 32$. This corroborates the theoretical result in Theorem 3.1. However, as the batch size m further increases to $m = 1,024$ close to $n = 1,500$, SGD starts to fail like GD and AM. Again, this phenomenon is explained by the fact that the noisy gradient in SGD avoids saddle points and local minima efficiently [23, 8].

For the Gaussian and uniform covariates, Figure 4c and Figure 4d illustrate the comparison in a high-sample regime, where the number of samples is twice larger than that for Figure 4a and Figure 4b, respectively. In this case, both GD and AM converge to a smaller error than SGD. Moreover, AM converges faster than the other algorithms in the run time,



(a) Gaussian distribution (b) Uniform distribution

Figure 5: Convergence of estimators for max-affine regression under additive white Gaussian noise of variance $\sigma^2 = 0.01$ ($k = 3$, $d = 500$, and $n = 8,000$).

which is explained by the following two reasons. First, as discussed in Section 2, AM converges faster than GD and SGD in the iteration count with a smaller constant for linear convergence. Second, due to the small ambient dimension ($d = 50$), the gain in the per-iteration cost of SGD $O(kmd)$ over that of AM $O(knd^2)$ is not significant.

Lastly, Figure 5, compares the convergence of the estimators in the presence of noise when d, k, and n are set as in Figure 1. On one hand, SGD converges faster than AM with a significantly lower per-iteration cost $O(kmd)$ than $O(knd^2)$ due to the large ambient dimension ($d = 500$) and small batch size ($m = 512$ compared to $n = 8,000$). On the other hand, SGD yields a larger error than the other two estimators. The estimation error bound of SGD by Theorem 3.1 behaves similarly in this case.

5. Discussion. We have established local convergence analysis of GD and SGD for max-affine regression under a relaxed covariate model with σ-sub-Gaussian noise. The covariate distribution characterized by the sub-Gaussianity and the anti-concentration generalizes beyond the standard Gaussian model. It has been shown that suitably initialized GD and SGD converge linearly below a non-asymptotic error bound, which is comparable to the analogous result on AM. Notably, when applied to noiseless max-affine regression, SGD empirically outperforms GD and AM in both sample complexity and convergence speed.

Under a special case of the Gaussian covariate model, the spectral method by Ghosh et al. [14] can provide the desired initial estimate. It is of great interest to extend their theory on the spectral method to the relaxed covariate model. Moreover, the extension of the theoretical result on GD and SGD to robust regression, where a subset of samples is corrupted as outliers, is also an intriguing future direction.

Acknowledgement. The authors thank Sohail Bahmani for the helpful discussions.

REFERENCES

[1] G. Balázs, Convex regression: theory, practice, and applications, (2016).
[2] G. Balázs, Adaptively partitioning max-affine estimators for convex regression, in International Conference on Artificial Intelligence and Statistics, PMLR, 2022, pp. 860–874.
[3] G. Balázs, A. György, and C. Szepesvári, Near-optimal max-affine estimators for convex regression, in Artificial Intelligence and Statistics, PMLR, 2015, pp. 56–64.
[4] L. Bottou, F. E. Curtis, and J. Nocedal, Optimization methods for large-scale machine learning, SIAM review, 60 (2018), pp. 223–311.
[5] S. Boyd, S. P. Boyd, and L. Vandenberghe, Convex optimization, Cambridge university press, 2004.
[6] F. H. Clarke, Generalized gradients and applications, Transactions of the American Mathematical Society, 205 (1975), pp. 247–262.
[7] K. Crammer and Y. Singer, On the algorithmic implementation of multiclass kernel-based vector machines, Journal of machine learning research, 2 (2001), pp. 265–292.
[8] H. Daneshmand, J. Kohler, A. Lucchi, and T. Hofmann, Escaping saddles with stochastic gradients, in International Conference on Machine Learning, PMLR, 2018, pp. 1155–1164.
[9] A. Daniely, S. Sabato, and S. Shwartz, Multiclass learning approaches: A theoretical comparison with implications, Advances in Neural Information Processing Systems, 25 (2012).
[10] A. Engel, Statistical mechanics of learning, Cambridge University Press, 2001.
[11] C. Finn, P. Abbeel, and S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in International conference on machine learning, PMLR, 2017, pp. 1126–1135.
[12] A. Ghosh, A. Pananjady, A. Guntuboyina, and K. Ramchandran, Max-affine regression: Provable, tractable, and near-optimal statistical estimation, arXiv preprint arXiv:1906.09255, (2019).
[13] A. Ghosh, A. Pananjady, A. Guntuboyina, and K. Ramchandran, Max-affine regression with universal parameter estimation for small-ball designs, in 2020 IEEE International Symposium on Information Theory (ISIT), IEEE, 2020, pp. 2706–2710.
[14] A. Ghosh, A. Pananjady, A. Guntuboyina, and K. Ramchandran, Max-affine regression: Parameter estimation for gaussian designs, IEEE Transactions on Information Theory, (2021).
[15] S. Goldt, M. Advani, A. M. Saxe, F. Krzakala, and L. Zdeborová, Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup, Advances in neural information processing systems, 32 (2019).
[16] I. Goodfellow, Y. Bengio, and A. Courville, Deep learning, MIT press, 2016.
[17] A. Guntuboyina and B. Sen, Covering numbers for convex functions, IEEE Transactions on Information Theory, 59 (2012), pp. 1957–1965.
[18] Q. Han and J. A. Wellner, Multivariate convex regression: global risk bounds and adaptation, arXiv preprint arXiv:1601.06844, (2016).
[19] L. A. Hannah and D. B. Dunson, Multivariate convex regression with adaptive partitioning, The Journal of Machine Learning Research, 14 (2013), pp. 3261–3294.
[20] N. J. Harvey, C. Liaw, Y. Plan, and S. Randhawa, Tight analyses for non-smooth stochastic gradient descent, in Conference on Learning Theory, PMLR, 2019, pp. 1579–1613.
[21] J. Hiriart-Urruty, New concepts in nondifferentiable programming, Bull. Soc. Math. France, 60 (1979), pp. 57–85.
[22] T. Hu, Z. Shang, and G. Cheng, Sharp rate of convergence for deep neural network classifiers under the teacher-student setting, arXiv preprint arXiv:2001.06892, (2020).
[23] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, How to escape saddle points efficiently, in International conference on machine learning, PMLR, 2017, pp. 1724–1732.
[24] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, and R. Cummings, Advances and open problems in federated learning, Foundations and Trends® in Machine Learning, 14 (2021), pp. 1–210.
[25] S. M. M. Kalan, M. Soltanolkotabi, and A. S. Avestimehr, Fitting relus via sgd and quantized sgd, in 2019 IEEE International Symposium on Information Theory (ISIT), IEEE, 2019, pp. 2469–2473.
[26] S. Kim, S. Bahmani, and K. Lee, Max-linear regression by scalable and guaranteed convex programming, arXiv preprint arXiv:2103.07020, (2021).
[27] G. Kur, Y. Dagan, and A. Rakhlin, Optimality of maximum likelihood for log-concave density estimation and bounded convex regression, arXiv preprint arXiv:1903.05315, (2019).
[28] E. Lim and P. W. Glynn, Consistency of multidimensional convex regression, Operations Research, 60

(2012), pp. 196–208.
[29] C. Mace and A. Coolen, Statistical mechanical analysis of the dynamics of learning in perceptrons, Statistics and Computing, 8 (1998), pp. 55–88.
[30] A. Magnani and S. P. Boyd, Convex piecewise-linear fitting, Optimization and Engineering, 10 (2009), pp. 1–17.
[31] J. Morgenstern and T. Roughgarden, Learning simple auctions, in Conference on Learning Theory, PMLR, 2016, pp. 1298–1318.
[32] D. Needell, R. Ward, and N. Srebro, Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm, Advances in neural information processing systems, 27 (2014).
[33] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, Robust stochastic approximation approach to stochastic programming, SIAM Journal on optimization, 19 (2009), pp. 1574–1609.
[34] A. Rubinstein and S. M. Weinberg, Simple mechanisms for a subadditive buyer and applications to revenue monotonicity, ACM Transactions on Economics and Computation (TEAC), 6 (2018), pp. 1–25.
[35] A. Siahkamari, D. A. E. Acar, C. Liao, K. L. Geyer, V. Saligrama, and B. Kulis, Faster algorithms for learning convex functions, in International Conference on Machine Learning, PMLR, 2022, pp. 20176–20194.
[36] A. Siahkamari, V. Saligrama, D. Castanon, and B. Kulis, Learning Bregman divergences, arXiv preprint arXiv:1905.11545, (2019).
[37] A. Siahkamari, X. Xia, V. Saligrama, D. Castañón, and B. Kulis, Learning to approximate a bregman divergence, Advances in Neural Information Processing Systems, 33 (2020), pp. 3603–3612.
[38] M. Soltanolkotabi, Learning relus via gradient descent, Advances in neural information processing systems, 30 (2017).
[39] R. S. Sutton and A. G. Barto, Reinforcement learning: An introduction, MIT press, 2018.
[40] Y. S. Tan and R. Vershynin, Online stochastic gradient descent with arbitrary initialization solves non-smooth, non-convex phase retrieval, arXiv preprint arXiv:1910.12837, (2019).
[41] Y. S. Tan and R. Vershynin, Phase retrieval via randomized kaczmarz: theoretical guarantees, Information and Inference: A Journal of the IMA, 8 (2019), pp. 97–123.
[42] A. Toriello and J. P. Vielma, Fitting piecewise linear continuous functions, European Journal of Operational Research, 219 (2012), pp. 86–95.
[43] G. Vardi, G. Yehudai, and O. Shamir, Learning a single neuron with bias using gradient descent, Advances in Neural Information Processing Systems, 34 (2021).
[44] R. Vershynin, High-dimensional probability: An introduction with applications in data science, vol. 47, Cambridge university press, 2018.
[45] G. Wang, G. B. Giannakis, Y. Saad, and J. Chen, Phase retrieval via reweighted amplitude flow, IEEE Transactions on Signal Processing, 66 (2018), pp. 2818–2833.
[46] G. Yehudai and S. Ohad, Learning a single neuron with gradient methods, in Conference on Learning Theory, PMLR, 2020, pp. 3756–3786.
[47] H. Zhang, Y. Zhou, Y. Liang, and Y. Chi, A nonconvex approach for phase retrieval: Reshaped wirtinger flow and incremental algorithms, Journal of Machine Learning Research, 18 (2017).
[48] X. Zhang, Y. Yu, L. Wang, and Q. Gu, Learning one-hidden-layer relu networks via gradient descent, in The 22nd international conference on artificial intelligence and statistics, PMLR, 2019, pp. 1524–1534.
[49] K. Zhong, Z. Song, P. Jain, P. L. Bartlett, and I. S. Dhillon, Recovery guarantees for one-hidden-layer neural networks, in International conference on machine learning, PMLR, 2017, pp. 4140–4149.

SUPPLEMENTARY MATERIALS: Max-Affine Regression via first-order methods*

Seonho Kim† and Kiryung Lee†

SM1. Tools. This section collects a set of standard results on concentration inequalities, which will be used in the proofs of Theorem 2.1. The following lemma provides the concentration of extreme singular values of sub-Gaussian matrices.

Lemma SM1.1 ([SM11, Theorem 4.6.1]). *Let $\{\boldsymbol{x}_i\}_{i=1}^n$ be independent isotropic η-sub-Gaussian random vectors in $\mathbb{R}^d$. Then there exists an absolute constant $C > 0$ such that*

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top - \boldsymbol{I}_p\right\| > \eta^2 \max(\epsilon, \epsilon^2)\right) \le \delta \quad \text{where} \quad \epsilon = \sqrt{\frac{C(d + \log(2/\delta))}{n}}.$$

Remark SM1.2. It has been shown that Lemma SM1.1 continues to hold when $\boldsymbol{x}_i$ is substituted by $\boldsymbol{\xi} = [\boldsymbol{x}_i;\ 1]$ [SM3]. Indeed, multiplying a random sign to the last coordinate of $\boldsymbol{\xi}_i$ does not modify the outer product $\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top$ whereas $\boldsymbol{\xi}_i$ remains a sub-Gaussian vector.

Furthermore, we also use the results from the standard Vapnik–Chervonenkis (VC) theory stated in the following lemmas.

Lemma SM1.3 ([SM10, Theorem 2]). *Let $\mathcal{V}$ be a collection of subsets of a set $\mathcal{X}$ and $\{\boldsymbol{x}_i\}_{i=1}^n$ be n independent copies of a random variable $\boldsymbol{x} \in \mathcal{X}$. Then it holds for all $\epsilon > 0$ and $n \ge 2/\epsilon^2$ that*

$$\mathbb{P}\left(\sup_{V\in\mathcal{V}}\left|\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i\in V\}} - \mathbb{P}(\boldsymbol{x}\in V)\right| \ge \epsilon\right) \le 4\Pi_{\mathcal{V}}(2n)\exp(-n\epsilon^2/16),$$

where $\Pi_{\mathcal{V}}(n)$ denotes the growth function defined by

$$\Pi_{\mathcal{V}}(n) := \max_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\in\mathcal{X}} \left|\left\{\left(\mathbb{1}_{\{\boldsymbol{x}_1\in V\}}, \ldots, \mathbb{1}_{\{\boldsymbol{x}_n\in V\}}\right) : V \in \mathcal{V}\right\}\right|.$$

Lemma SM1.4 ([SM8, Corollary 3.18]). *Let $\mathcal{V}$ be a collection of subsets having VC dimension d. Then, for all $n \ge d$, the growth function of $\mathcal{V}$ is upper-bounded by*

$$\Pi_{\mathcal{V}}(n) \le \left(\frac{en}{d}\right)^d.$$

The VC dimension of the k-fold intersection has been known in the literature (e.g. see [SM1]). We will use the following lemma for the result for the intersection of size two. Since it was given as an exercise in [SM8], we provide a proof for the sake of completeness.

---

*Submitted to the editors on March 17, 2024.

Funding: This work was supported in part by NSF CAREER Award CCF 19-43201.

†The Ohio State University, Columbus, Ohio (kim.7604@osu.edu, kiryung@ece.osu.edu).

Lemma SM1.5 ([SM8, Equation (3.53)]). *Let $\mathcal{V}$ and $\mathcal{W}$ be collections of subsets of a common set. Then their intersection given by $\mathcal{V} \cap \mathcal{W} := \{V \cap W : V \in \mathcal{V}, W \in \mathcal{W}\}$ satisfies that*

$$\Pi_{\mathcal{V}\cap\mathcal{W}}(n) \leq \Pi_{\mathcal{V}}(n)\Pi_{\mathcal{W}}(n), \quad \forall n \in \mathbb{N}.$$

Proof. For any $V \cap W \in \mathcal{V} \cap \mathcal{W}$, we have

$$\left(\mathbb{1}_{\{\boldsymbol{x}_1 \in V\cap W\}}, \dots, \mathbb{1}_{\{\boldsymbol{x}_n \in V\cap W\}}\right) = \left(\mathbb{1}_{\{\boldsymbol{x}_1 \in V\}}, \dots, \mathbb{1}_{\{\boldsymbol{x}_n \in V\}}\right) \odot \left(\mathbb{1}_{\{\boldsymbol{x}_1 \in W\}}, \dots, \mathbb{1}_{\{\boldsymbol{x}_n \in W\}}\right),$$

where $\odot$ denotes the pointwise product. Therefore, the claim follows from the definition of the growth function. ■

Lemma SM1.6. *Let $\mathcal{P}_k$ be the collection of all polytopes constructed by the intersection of k half spaces in $\mathbb{R}^d$. Then the growth function of $\mathcal{P}_k$ satisfies*

$$\Pi_{\mathcal{P}_k}(n) \leq \left(\frac{en}{d+1}\right)^{k(d+1)}. \tag{SM1.1}$$

Proof. Let $\mathcal{H}_j$ be the collection of all half spaces in $\mathbb{R}^d$ for $j \in [k]$. Then, by the construction of $\mathcal{P}_k$, we have $\mathcal{P}_k = \cap_{j=1}^k \mathcal{H}_j$. Therefore, by inductive application of Lemma SM1.5, the growth function of $\mathcal{P}_k$ satisfies

$$\Pi_{\mathcal{P}_k}(n) \leq \prod_{j=1}^{k} \Pi_{\mathcal{H}_j}(n). \tag{SM1.2}$$

Furthermore, since the VC dimensions of half spaces in $\mathbb{R}^d$ is $d+1$ (e.g. see [SM8, Section 3]), Lemma SM1.4 implies

$$\Pi_{\mathcal{H}_j}(n) \leq \left(\frac{en}{d+1}\right)^{d+1}, \quad \forall j \in [k]. \tag{SM1.3}$$

The assertion is obtained by plugging in (SM1.3) into (SM1.2). ■

Finally, the following corollary is a direct consequence of Lemmas SM1.3, SM1.4, and SM1.5.

Corollary SM1.7. *Let $\delta \in (0,1)$ and $\mathcal{P}_k$ be the collection of all polytopes constructed by the intersection of k half-spaces in $\mathbb{R}^d$. Suppose that $\{\boldsymbol{x}_i\}_{i=1}^n$ are independent copies of a random vector $\boldsymbol{x} \in \mathbb{R}^d$. Then it holds with probability at least $1-\delta$ that*

$$\sup_{Z\in\mathcal{P}_k} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{x}_i \in Z\}} - \mathbb{P}(\boldsymbol{x} \in Z) \right| \leq 4\sqrt{\frac{\log(4/\delta) + 2k(d+1)\log(2en/(d+1))}{n}}. \tag{SM1.4}$$

SM2. Supporting lemmas. In this section, we list lemmas to prove Theorem 2.1. These lemmas are borrowed from [SM9] and [SM3]. We improve on a subset of these results derived with a streamlined proof.

SM2.1. Worst-case extreme eigenvalues of partial sum of outer products of covariates.

A partial sum of the outer products of covariates, $\sum_{i\in\mathcal{I}} \boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top$ appears frequently in the proof. The summation indices in $\mathcal{I}$ often depend on covariates. The following lemma by Tan and Vershynin [SM9] provides a tail bound on the worst-case largest eigenvalue of $\sum_{i\in\mathcal{I}} \boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top$ when the cardinality of $\mathcal{I}$ is bounded from above.

Lemma SM2.1 ([SM9, Theorem 5.7]). *Let $\delta \in (0, 1/e)$, $\alpha \in (0, 1)$, and $\boldsymbol{\xi}_i = [\boldsymbol{x}_i, 1] \in \mathbb{R}^{d+1}$ for $i \in [n]$. Suppose that Assumption 1.1 holds. Then it holds with probability at least $1 - \delta$ that*

$$\sup_{\mathcal{I}:|\mathcal{I}|\le \alpha n} \lambda_1\left(\sum_{i\in\mathcal{I}} \boldsymbol{\xi}_i\boldsymbol{\xi}_i^\intercal\right) \le C_4(\eta^2 \vee 1)\sqrt{\alpha} n$$

for some absolute constant $C_4 > 0$, provided

$$n \ge \left(d \vee \frac{\log(1/\delta)}{\alpha}\right). \tag{SM2.1}$$

Remark SM2.2. In the original result, Tan and Vershynin assumed that $\{\boldsymbol{\xi}_i\}_{i=1}^n$ are isotropic η-sub-Gaussian random vectors [SM9, Theorem 5.7]. Later, Ghosh et al. [SM3] showed that the result also applies to the setting in Lemma SM2.1 through the following argument. The outer product $\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top$ remains the same as one multiplies a random sign to the last entry of $\boldsymbol{\xi}_i$ which makes the random vector $\widetilde{\eta}$-sub-Gaussian with $\widetilde{\eta} = \max(\eta, 1)$.

Moreover, Ghosh et al. also derived analogous lower tail bound on the smallest eigenvalue when the index set $\mathcal{I}$ exceeds a threshold [SM3, Lemma 7]. Their proof strategy adopted an epsilon-net approximation and a union bound argument. Our lemma below, derived by using the small-ball method [SM6], provides a streamlined proof and a sharper bound.

Lemma SM2.3. *Let $\alpha, \delta \in (0, 1)$ and $\boldsymbol{\xi}_i = [\boldsymbol{x}_i, 1] \in \mathbb{R}^{d+1}$ for $i \in [n]$. Suppose that Assumption 1.2 holds. Then there exists an absolute constant $C > 0$ such that if*

$$n \ge C\alpha^{-2}(d \log(n/d) \vee \log(1/\delta)) \tag{SM2.2}$$

then it holds with probability at least $1 - \delta$ that

$$\inf_{\mathcal{I}\subset[n]:|\mathcal{I}|\ge\alpha n} \lambda_{d+1}\left(\sum_{i\in\mathcal{I}} \boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top\right) \ge \frac{2n}{\gamma}\left(\frac{\alpha}{4}\right)^{1+\zeta^{-1}}. \tag{SM2.3}$$

We compare Lemma SM2.3 to the previous result by Ghosh et al. [SM3, Lemma 7] when the parameter γ is treated as a fixed constant. They demonstrated that the worst-case minimum eigenvalue in the left-hand side of (SM2.3) satisfies $\Omega(n\alpha^{1+2\zeta^{-1}})$ if $n \ge \alpha^{-1}\max(4p, \zeta^{-1}(d+1))$. On one hand, their requirement in the sample complexity is less stringent than that in (SM2.2). On the other hand, the lower bound in (SM2.3) is tighter than theirs by a factor of $\alpha^{\zeta^{-1}}$. When these two results are applied to derive Theorem 2.1 with α substituted by $\pi_{\min}$, the resulting sample complexity $\widetilde{O}(\pi_{\min}^{-4(1+\zeta^{-1})}d)$ by Lemma SM2.3 is smaller than $\widetilde{O}(\pi_{\min}^{-4(1+2\zeta^{-1})}d)$ by [SM3, Lemma 7]. The gain due to Lemma SM2.3 is $\pi_{\min}^{-4\zeta^{-1}}$, which is no less than $k^{4\zeta^{-1}}$. For example, if the covariates are Gaussian $\zeta = 1/2$, then the gain is k^8.

Proof. Let $T > 0$ be an arbitrarily fixed threshold. If

$$N(\boldsymbol{v}) := \sum_{i=1}^{n} \mathbb{1}_{\{\langle \boldsymbol{\xi}_i, \boldsymbol{v}\rangle^2 > T\}} > n - \frac{\alpha n}{2} \tag{SM2.4}$$

then it follows that

$$\frac{1}{n} \sum_{i \in \mathcal{I}} \langle \boldsymbol{\xi}_i, \boldsymbol{v}\rangle^2 \geq \frac{\alpha T}{2}, \quad \forall \mathcal{I} \subset [n] : |\mathcal{I}| \geq \alpha n.$$

Therefore, it suffices to show that (SM2.4) holds for all $\boldsymbol{v} \in \mathbb{S}^d$ with probability $1-\delta$. Let $\mathcal{H}$ denote the collection of half-spaces in $\mathbb{R}^d$ given by $\{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{x}^\top \boldsymbol{u} > \sqrt{T} - w\}$ for all $\boldsymbol{v} = [\boldsymbol{u};\ w] \in \mathbb{S}^d$. Since the VC dimension of all half-spaces in $\mathbb{R}^d$ is at most $d+1$, by Lemmas SM1.3 and SM1.4, it holds with probability at least $1-\delta/2$ that

$$\frac{1}{n} N(\boldsymbol{v}) \geq \frac{1}{n} \mathbb{E} N(\boldsymbol{v}) - C' \sqrt{\frac{d \log(n/d) + \log(1/\delta)}{n}}, \quad \forall \boldsymbol{v} \in \mathbb{S}^d, \tag{SM2.5}$$

where $C' > 0$ is an absolute constant.

Moreover, it follows from Assumption 1.2 that

$$\frac{1}{n} \mathbb{E} N(\boldsymbol{v}) = \mathbb{P}\left(|\langle \boldsymbol{x}, \boldsymbol{u}\rangle + w|^2 > T\right) \geq 1 - (T\gamma)^\zeta. \tag{SM2.6}$$

By plugging in (SM2.6) into (SM2.5), we obtain that

$$\frac{1}{n} N(\boldsymbol{v}) \geq 1 - (T\gamma)^\zeta - C' \sqrt{\frac{d \log(n/d) + \log(1/\delta)}{n}}, \quad \forall \boldsymbol{w} \in \mathbb{S}^d.$$

Then (SM2.4) is satisfied for all $\boldsymbol{v} \in \mathbb{S}^d$ when $T = \frac{1}{\gamma}\left(\frac{\alpha}{4}\right)^{\zeta^{-1}}$ and $C = (4C')^2$. This completes the proof. ∎

SM2.2. Local estimates. In this section, we present local tail bounds which arise in the proof of the main result. The following lemma, obtained as a direct consequence of the triangle inequality and the definition of κ in (2.11), provides a basic inequality that will be used frequently throughout this section.

Lemma SM2.4. *Suppose that $\boldsymbol{\beta} \in \mathcal{N}(\boldsymbol{\beta}^\star)$, where $\mathcal{N}(\boldsymbol{\beta}^\star)$ is defined as in (2.12). Then we have*

$$\|(\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j'}) - (\boldsymbol{\beta}_j^\star - \boldsymbol{\beta}_{j'}^\star)\|_2 \leq 2\rho \|(\boldsymbol{\beta}_j^\star - \boldsymbol{\beta}_{j'}^\star)_{1:d}\|_2, \quad \forall j \neq j' \in [k].$$

Proof. Since $\boldsymbol{\beta} \in \mathcal{N}(\boldsymbol{\beta}^\star)$, by the triangle inequality, we have

$$\|(\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j'}) - (\boldsymbol{\beta}_j^\star - \boldsymbol{\beta}_{j'}^\star)\|_2 \leq \|\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^\star\|_2 + \|\boldsymbol{\beta}_{j'} - \boldsymbol{\beta}_{j'}^\star\|_2 \leq 2\kappa\rho, \quad \forall j, j' \in [k].$$

Furthermore, it follows from the definition of κ in (2.11) that

$$\kappa \leq \|(\boldsymbol{\beta}_j^\star - \boldsymbol{\beta}_{j'}^\star)_{1:d}\|_2, \quad \forall j \neq j' \in [k].$$

Then the assertion follows. ∎

We also use the following lemma by Ghosh et al. [SM3], which is a consequence of Assumptions 1.1 and 1.2 respectively for the sub-Gaussianity and anti-concentration.

Lemma SM2.5 ([SM3, Lemma 17]). *Suppose that $\boldsymbol{x} \in \mathbb{R}^d$ satisfies Assumptions 1.1 and 1.2. If*

$$\|\boldsymbol{v} - \boldsymbol{v}^\star\|_2 \leq \frac{1}{2}\|(\boldsymbol{v}^\star)_{1:d}\|_2,$$

then

$$\mathbb{P}\left(\langle[\boldsymbol{x};1], \boldsymbol{v}^\star\rangle^2 \leq \langle[\boldsymbol{x};1], \boldsymbol{v} - \boldsymbol{v}^\star\rangle^2\right) \lesssim \left(\left(\frac{\|\boldsymbol{v} - \boldsymbol{v}^\star\|_2}{\|(\boldsymbol{v}^\star)_{1:d}\|_2}\right)^2 \cdot \log\left(\frac{2\|(\boldsymbol{v}^\star)_{1:d}\|_2}{\|\boldsymbol{v} - \boldsymbol{v}^\star\|_2}\right)\right)^\zeta.$$

Intuitively, when the parameter vector $\boldsymbol{\beta}$ belongs to a small neighborhood of the ground-truth, the partition sets $(\mathcal{C}_j)_{j=1}^k$ by $\boldsymbol{\beta}$ and $\left(\mathcal{C}_j^\star\right)_{j=1}^k$ by the ground-truth $\boldsymbol{\beta}^\star$ will be similar. The next lemmas quantify the empirical measure on the event of $\boldsymbol{x} \in \mathcal{C}_j \cap \mathcal{C}_{j'}^\star$ for distinct indices j and j', and quadratic forms given as a partial summation indexed by the indicator functions on this event.

Lemma SM2.6. *Let $(\mathcal{C}_j)_{j=1}^k$ and $\left(\mathcal{C}_j^\star\right)_{j=1}^k$ be defined as in (2.4) and (2.10) respectively by $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^\star$. Furthermore, let $\pi_{\min}$ be defined as in (2.9) by $\boldsymbol{\beta}^\star$. Suppose that $\boldsymbol{x} \in \mathbb{R}^d$ and $\{\boldsymbol{x}_i\}_{i=1}^n$ satisfy Assumptions 1.1 and 1.2, and that the parameter ρ of $\mathcal{N}(\boldsymbol{\beta}^\star)$ in (2.12) satisfies (2.13) for some numerical constant $R > 0$. Then there exists an absolute constant C such that if*

$$n \geq C\pi_{\min}^{-2} \cdot (kd\log(n/d) \vee \log(1/\delta)) \tag{SM2.7}$$

then with probability at least $1 - \delta$

$$\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j \cap \mathcal{C}_j^\star\}} \geq \frac{\pi_{\min}}{4} \tag{SM2.8}$$

holds for all $j \in [k]$, $\boldsymbol{\beta} \in \mathcal{N}(\boldsymbol{\beta}^\star)$, and $\boldsymbol{\beta}^\star \in \mathbb{R}^{d+1}$.

Proof. Note that the left-hand side of (SM2.8) is an empirical measure on the event $\boldsymbol{x} \in \mathcal{C}_j \cap \mathcal{C}_j^\star$. We first derive a lower bound on its expectation, which is written as

$$\begin{aligned}\mathbb{P}\left(\boldsymbol{x} \in \mathcal{C}_j, \boldsymbol{x} \in \mathcal{C}_j^\star\right) &= \mathbb{P}\left(\boldsymbol{x} \in \mathcal{C}_j | \boldsymbol{x} \in \mathcal{C}_j^\star\right) \cdot \mathbb{P}\left(\boldsymbol{x} \in \mathcal{C}_j^\star\right) \\ &= \left(1 - \mathbb{P}\left(\boldsymbol{x} \notin \mathcal{C}_j | \boldsymbol{x} \in \mathcal{C}_j^\star\right)\right) \cdot \mathbb{P}\left(\boldsymbol{x} \in \mathcal{C}_j^\star\right).\end{aligned} \tag{SM2.9}$$

Then, by the construction of $(\mathcal{C}_j)_{j=1}^k$ in (2.4), we have

$$
\begin{aligned}
&\mathbb{P}\left(\boldsymbol{x} \notin \mathcal{C}_j | \boldsymbol{x} \in \mathcal{C}_j^\star\right) \\
&= \frac{\mathbb{P}(\boldsymbol{x} \notin \mathcal{C}_j, \boldsymbol{x} \in \mathcal{C}_j^\star)}{\mathbb{P}(\boldsymbol{x} \in \mathcal{C}_j^\star)} \\
&\leq \frac{1}{\mathbb{P}(\boldsymbol{x} \in \mathcal{C}_j^\star)} \sum_{j' \neq j} \mathbb{P}\left(\langle [\boldsymbol{x};\ 1], \boldsymbol{\beta}_{j'}\rangle \geq \langle [\boldsymbol{x};\ 1], \boldsymbol{\beta}_j\rangle, \langle [\boldsymbol{x};\ 1], \boldsymbol{\beta}_j^\star\rangle \geq \langle [\boldsymbol{x};\ 1], \boldsymbol{\beta}_{j'}^\star\rangle\right) \\
&\leq \frac{1}{\mathbb{P}(\boldsymbol{x} \in \mathcal{C}_j^\star)} \sum_{j' \neq j} \mathbb{P}\left(\langle [\boldsymbol{x};\ 1], \boldsymbol{v}_{j,j'}\rangle \langle [\boldsymbol{x};\ 1], \boldsymbol{v}_{j,j'}^\star\rangle \leq 0\right) \\
&\leq \frac{1}{\mathbb{P}(\boldsymbol{x} \in \mathcal{C}_j^\star)} \sum_{j' \neq j} \mathbb{P}\left(\langle [\boldsymbol{x}; 1], \boldsymbol{v}_{j,j'}^\star\rangle^2 \leq \langle [\boldsymbol{x}; 1], \boldsymbol{v}_{j,j'} - \boldsymbol{v}_{j,j'}^\star\rangle^2\right),
\end{aligned}
$$

where the second inequality holds since $\boldsymbol{v}_{j,j'} = \boldsymbol{\beta}_j - \boldsymbol{\beta}_{j'}$ and $\boldsymbol{v}_{j,j'}^\star = \boldsymbol{\beta}_j^\star - \boldsymbol{\beta}_{j'}^\star$, and the last inequality follows from the fact that $ab \leq 0$ implies $|b| \leq |a - b|$ for $a, b \in \mathbb{R}$. Recall that $\boldsymbol{\beta} \in \mathcal{N}(\boldsymbol{\beta}^\star)$ implies $\|\boldsymbol{v}_{j,j'} - \boldsymbol{v}_{j,j'}^\star\|_2 \leq 2\rho\|(\boldsymbol{v}_{j,j'}^\star)_{1:d}\|_2$ due to Lemma SM2.4. Furthermore, one can choose the numerical constant $R > 0$ in (2.13) sufficiently small (but independent of k and p) so that $2\rho \leq 0.1$. Then it follows that

$$
\begin{aligned}
\mathbb{P}(\boldsymbol{x} \notin \mathcal{C}_{j'} \,|\, \boldsymbol{x} \in \mathcal{C}_{j'}^\star) &\overset{\text{(i)}}{\lesssim} \frac{k}{\mathbb{P}(\boldsymbol{x} \in \mathcal{C}_j^\star)} \left(\frac{\|\boldsymbol{v}_{j,j'} - \boldsymbol{v}_{j,j'}^\star\|_2^2}{\|(\boldsymbol{v}_{j,j'}^\star)_{1:d}\|_2^2} \log\left(\frac{2\|(\boldsymbol{v}_{j,j'}^\star)_{1:d}\|_2}{\|\boldsymbol{v}_{j,j'} - \boldsymbol{v}_{j,j'}^\star\|_2}\right)\right)^\zeta \\
&\overset{\text{(ii)}}{\leq} \frac{k}{\mathbb{P}(\boldsymbol{x} \in \mathcal{C}_j^\star)} \left((2\rho)^2 \log\left(\frac{1}{\rho}\right)\right)^\zeta \\
&\overset{\text{(iii)}}{\leq} \frac{k}{\pi_{\min}} \left(\frac{R^2 \pi_{\min}^{2\zeta^{-1}(1+\zeta^{-1})}}{k^{2\zeta^{-1}}}\right)^\zeta \\
&\leq \frac{R^{2\zeta} \pi_{\min}^{1+2\zeta^{-1}}}{k},
\end{aligned}
\tag{SM2.10}
$$

where (i) follows from Lemma SM2.5; (ii) holds since $a \log^{1/2}(2/a)$ is monotone increasing for $a \in (0, 1]$; (iii) follows from the fact that $a \leq \frac{b}{2} \log^{-1/2}(1/b)$ implies $a \log^{1/2}(2/a) \leq b$ for $b \in (0, 0.1]$. Since $\pi_{\min} \leq \frac{1}{k}$, once again $R > 0$ can be made sufficiently small so that the right-hand side of (SM2.10) is at most $\frac{1}{2}$. Then plugging in this upper bound by (SM2.10) into (SM2.9) yields

$$
\mathbb{P}(\boldsymbol{x} \in \mathcal{C}_{j'} \cap \mathcal{C}_{j'}^\star) \geq \frac{1}{2} \cdot \mathbb{P}(\boldsymbol{x} \in \mathcal{C}_{j'}^\star). \tag{SM2.11}
$$

It remains to show the concentration of the left-hand side of (SM2.8) around the expectation. Recall that $\mathcal{C}_j$ and $\mathcal{C}_j^\star$ are constructed as the intersection of at most k half-spaces. Then $\mathcal{C}_j \cap \mathcal{C}_j^\star$ belongs to the set $\mathcal{P}_{2k}$ defined in Lemma SM1.6 and, hence, we have

$$
\sup_{\substack{j \in [k], \boldsymbol{\beta} \in \mathcal{N}(\boldsymbol{\beta}^\star) \\ \boldsymbol{\beta}^\star \in \mathbb{R}^{d+1}}} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j \cap \mathcal{C}_j^\star\}} - \mathbb{P}(\boldsymbol{x} \in \mathcal{C}_j \cap \mathcal{C}_j^\star) \right| \leq \sup_{\mathcal{Z} \in \mathcal{P}_{2k}} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{Z}\}} - \mathbb{P}(\boldsymbol{x} \in \mathcal{Z}) \right|.
$$

Therefore, it follows from Corollary SM1.7 that with probability at least $1-\delta$

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\cap\mathcal{C}_j^\star\}} \geq \mathbb{P}(\boldsymbol{x}\in\mathcal{C}_j\cap\mathcal{C}_j^\star) - 4\sqrt{\frac{\log(4/\delta)+2k(d+1)\log(2en/(d+1))}{n}} \tag{SM2.12}$$

holds for all $j\in[k]$, $\boldsymbol{\beta}\in\mathcal{N}(\boldsymbol{\beta}^\star)$, and $\boldsymbol{\beta}^\star\in\mathbb{R}^{d+1}$. The first summand in the right-hand side of (SM2.12) is bounded from below as in (SM2.11). Then choosing C in (SM2.7) large enough makes the second summand less than half of the lower bound in (SM2.11). This completes the proof. ∎

Next, the following lemma provides a slightly improved upper bound compared to the analogous previous result [SM3, Lemma 6]. Moreover, Lemma SM2.7 is derived by using the VC theory and provides a streamlined and shorter proof compared to previous work [SM3].

Lemma SM2.7. *Suppose that Assumptions 1.1 and 1.2 hold, and that ρ satisfies (2.13) for some numerical constant $R>0$. Let $\delta\in(0,1/e)$. There exists an absolute constant C such that if*

$$n \geq Ck^4\pi_{\min}^{-4(1+\zeta^{-1})}(\log(k/\delta)\vee d\log(n/d)) \tag{SM2.13}$$

then with probability at least $1-\delta$

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\cap\mathcal{C}_{j'}^\star\}}\langle[\boldsymbol{x}_i;\ 1],\boldsymbol{v}_{j,j'}^\star\rangle^2 \leq \frac{2}{5\gamma k}\left(\frac{\pi_{\min}}{16}\right)^{1+\zeta^{-1}}\|\boldsymbol{v}_{j,j'}-\boldsymbol{v}_{j,j'}^\star\|_2^2 \tag{SM2.14}$$

holds for all $j\in[k]$, $\boldsymbol{\beta}\in\mathcal{N}(\boldsymbol{\beta}^\star)$, and $\boldsymbol{\beta}^\star\in\mathbb{R}^{d+1}$ where $\boldsymbol{v}_{j,j'}=\boldsymbol{\beta}_j-\boldsymbol{\beta}_{j'}$ and $\boldsymbol{v}_{j,j'}^\star=\boldsymbol{\beta}_j^\star-\boldsymbol{\beta}_{j'}^\star$.

The previous result [SM3, Lemma 6] showed that with probability at least $1-\delta$ the left-hand side of (SM2.14) is bounded from above by $\widetilde{O}((\pi_{\min}^{1+\zeta^{-1}}/k)\log^{\zeta/2+1}(k/(\pi_{\min}{}^{1+\zeta^{-1}})))$ if $n\geq O(\max(p,\log(1/\delta)))$. In contrast, Lemma SM2.7 provides a smaller upper bound by a logarithmic factor at the cost of increased sample complexity. However, the condition in (SM2.13) is implied by another sufficient condition from another step of the analysis; hence, it does not affect the main result in Theorem 2.1.

Proof. By the definition of $(\mathcal{C}_j)_{j=1}^k$ in (2.4), it holds for any $j\neq j'$ that

$$\begin{aligned}\boldsymbol{x}_i\in\mathcal{C}_j\cap\mathcal{C}_{j'}^\star &\iff \langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j\rangle\geq\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_{j'}\rangle,\ \langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_{j'}^\star\rangle\geq\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j^\star\rangle\\ &\iff \langle\boldsymbol{\xi}_i,\boldsymbol{v}_{j,j'}\rangle\geq 0,\ \langle\boldsymbol{\xi}_i,\boldsymbol{v}_{j,j'}^\star\rangle\leq 0\\ &\implies \langle\boldsymbol{\xi}_i,\boldsymbol{v}_{j,j'}\rangle\langle\boldsymbol{\xi}_i,\boldsymbol{v}_{j,j'}^\star\rangle\leq 0.\end{aligned} \tag{SM2.15}$$

Furthermore, by Lemma SM2.4, every $\boldsymbol{\beta}\in\mathcal{N}(\boldsymbol{\beta}^\star)$ satisfies $\|\boldsymbol{v}_{j,j'}-\boldsymbol{v}_{j,j'}^\star\|_2\leq 2\rho\|(\boldsymbol{v}_{j,j'}^\star)_{1:d}\|_2$. Therefore, it suffices to show that with probability at least $1-\delta$

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{\langle\boldsymbol{\xi}_i,\boldsymbol{v}\rangle\langle\boldsymbol{\xi}_i,\boldsymbol{v}^\star\rangle\leq 0\}}\langle\boldsymbol{\xi}_i,\boldsymbol{v}^\star\rangle^2 \leq \frac{2}{5\gamma k}\left(\frac{\pi_{\min}}{16}\right)^{1+\zeta^{-1}}\|\boldsymbol{v}-\boldsymbol{v}^\star\|_2^2 \tag{SM2.16}$$

holds for all $(\boldsymbol{v}, \boldsymbol{v}^\star) \in \mathcal{M}$, where

$$\mathcal{M} := \{(\boldsymbol{v}, \boldsymbol{v}^\star) \in \mathbb{R}^{d+1} \times \mathbb{R}^{d+1} : \|\boldsymbol{v} - \boldsymbol{v}^\star\| \le 2\rho \|(\boldsymbol{v})_{1:d}\|_2\}.$$

Since $ab \le 0$ implies $|b| \le |a-b|$ for $a, b \in \mathbb{R}$, each summand in the left-hand side of (SM2.16) is upper-bounded by

$$\begin{aligned}\mathbb{1}_{\{\langle \boldsymbol{\xi}_i, \boldsymbol{v}\rangle\langle \boldsymbol{\xi}_i, \boldsymbol{v}^\star\rangle \le 0\}} \langle \boldsymbol{\xi}_i, \boldsymbol{v}^\star\rangle^2 &\le \mathbb{1}_{\{\langle \boldsymbol{\xi}_i, \boldsymbol{v}^\star\rangle^2 \le \langle \boldsymbol{\xi}_i, \boldsymbol{v}-\boldsymbol{v}^\star\rangle^2\}} \langle \boldsymbol{\xi}_i, \boldsymbol{v}^\star\rangle^2 \\ &\le \mathbb{1}_{\{\langle \boldsymbol{\xi}_i, \boldsymbol{v}^\star\rangle^2 \le \langle \boldsymbol{\xi}_i, \boldsymbol{v}-\boldsymbol{v}^\star\rangle^2\}} \langle \boldsymbol{\xi}_i, \boldsymbol{v}-\boldsymbol{v}^\star\rangle^2.\end{aligned}$$

Before we proceed to the next step, for brevity, we introduce a shorthand notation given by

$$\mathcal{S}_{\boldsymbol{v},\boldsymbol{v}^\star} := \{\boldsymbol{\xi} \in \mathbb{R}^{d+1} : \langle \boldsymbol{\xi}, \boldsymbol{v}-\boldsymbol{v}^\star\rangle^2 \ge \langle \boldsymbol{\xi}, \boldsymbol{v}^\star\rangle^2\}. \tag{SM2.17}$$

Then the left-hand side of (SM2.16) is bounded from above as

$$\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\langle \boldsymbol{\xi}_i, \boldsymbol{v}\rangle\langle \boldsymbol{\xi}_i, \boldsymbol{v}^\star\rangle \le 0\}} \langle \boldsymbol{\xi}_i, \boldsymbol{v}^\star\rangle^2 \le \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{\xi}_i \in \mathcal{S}_{\boldsymbol{v},\boldsymbol{v}^\star}\}} \langle \boldsymbol{\xi}_i, \boldsymbol{v}-\boldsymbol{v}^\star\rangle^2.$$

Next, we derive a tail bound on the empirical measure $\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{\xi}_i \in \mathcal{S}_{\boldsymbol{v},\boldsymbol{v}^\star}\}}$ on the event for $\boldsymbol{\xi} \in \mathcal{S}_{\boldsymbol{v},\boldsymbol{v}^\star}$. Let $\mathcal{P}_2$ denote the collection of all polytopes given by the intersections of two half-spaces. Then $\mathcal{S}_{\boldsymbol{v},\boldsymbol{v}^\star}$ belongs to $\mathcal{P}_2 \cup \mathcal{P}_2$. It follows from Lemma SM1.6 and [SM2, Theorem A] that

$$\Pi_{\mathcal{P}_2 \cup \mathcal{P}_2}(n) \le \left(\frac{en}{C'(d+1)}\right)^{C'(d+1)} \tag{SM2.18}$$

for some absolute constant C'. Therefore, by Lemma SM1.3 and (SM2.18), we obtain that

$$\sup_{(\boldsymbol{v},\boldsymbol{v}^\star)\in\mathcal{M}} \left|\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{\xi}_i \in \mathcal{S}_{\boldsymbol{v},\boldsymbol{v}^\star}\}} - \mathbb{P}(\boldsymbol{\xi} \in \mathcal{S}_{\boldsymbol{v},\boldsymbol{v}^\star})\right| \lesssim \sqrt{\frac{\log(1/\delta) + d\log(n/d)}{n}} \tag{SM2.19}$$

holds with probability at least $1 - \frac{\delta}{2}$.

Similar to (SM2.10), we obtain an upper bound on the probability by using Lemma SM2.5 as follows:

$$\begin{aligned}\sup_{(\boldsymbol{v},\boldsymbol{v}^\star)\in\mathcal{M}} \mathbb{P}(\boldsymbol{\xi} \in \mathcal{S}_{\boldsymbol{v},\boldsymbol{v}^\star}) &\le C_1\left((2\rho)^2 \log\left(\frac{1}{\rho}\right)\right)^\zeta \\ &\le C_1\left(\frac{R^2 \pi_{\min}^{2\zeta^{-1}(1+\zeta^{-1})}}{k^{2\zeta^{-1}}}\right)^\zeta \\ &\le \underbrace{\frac{C_1 R^{2\zeta} \pi_{\min}^{2+2\zeta^{-1}}}{k^2}}_{\alpha} \qquad \text{(SM2.20)}\end{aligned}$$

where $C_1 > 0$ is an absolute constant. By choosing the numerical constant $C > 0$ in (SM2.13) sufficiently large, we obtain from (SM2.19) and (SM2.20) that

$$\mathbb{P}\left(\sup_{(\boldsymbol{v},\boldsymbol{v}^\star)\in\mathcal{M}} \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{\xi}_i \in \mathcal{S}_{\boldsymbol{v},\boldsymbol{v}^\star}\}} > \frac{\alpha}{2} \right) \le \frac{\delta}{2}. \tag{SM2.21}$$

Furthermore, one can choose the numerical constant $R > 0$ small enough so that $\alpha \in (0,1)$. Then, since (SM2.13) and (2.13) imply (SM2.1), by Lemma SM2.1, it holds with probability at least $1-\delta/2$ that

$$\sup_{\mathcal{I}:|\mathcal{I}|\le\frac{\alpha n}{2}} \left\| \sum_{i\in\mathcal{I}} \boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top \right\| \lesssim (\eta^2 \vee 1)\sqrt{\alpha}n. \tag{SM2.22}$$

Finally, by combining the results in (SM2.21) and (SM2.22), we obtain that with probability at least $1-\delta$

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\langle\boldsymbol{\xi}_i,\boldsymbol{v}\rangle\langle\boldsymbol{\xi}_i,\boldsymbol{v}^\star\rangle\le 0\}}\langle\boldsymbol{\xi}_i,\boldsymbol{v}^\star\rangle^2 &\le \sup_{\mathcal{I}:|\mathcal{I}|\le\frac{\alpha n}{2}} \frac{1}{n}\sum_{i\in\mathcal{I}}\langle\boldsymbol{\xi}_i,\boldsymbol{v}-\boldsymbol{v}^\star\rangle^2 \\
&\le \sup_{\mathcal{I}:|\mathcal{I}|\le\frac{\alpha n}{2}} \left\| \frac{1}{n}\sum_{i\in\mathcal{I}} \boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top \right\| \cdot \|\boldsymbol{v}-\boldsymbol{v}^\star\|_2^2 \\
&\le C_2(\eta^2\vee 1)R^\zeta \left(\frac{\pi_{\min}^{(1+\zeta^{-1})}}{k} \right) \cdot \|\boldsymbol{v}-\boldsymbol{v}^\star\|_2^2
\end{aligned}$$

holds for all $(\boldsymbol{v},\boldsymbol{v}^\star)\in\mathcal{M}$, where C_2 is an absolute constant. By choosing $R > 0$ sufficiently small so that

$$C_2(\eta^2\vee 1)R^\zeta \le \frac{2}{5\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}},$$

we obtain the assertion in (SM2.16). ∎

SM3. Proof of Theorem 2.1.

The loss function $\ell(\boldsymbol{\beta})$ is decomposed as

$$\begin{aligned}
\ell(\boldsymbol{\beta}) &= \frac{1}{2n}\left(\max_{j\in[k]}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j\rangle - \max_{j\in[k]}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j^\star\rangle - z_i \right)^2 \\
&= \underbrace{\frac{1}{2n}\sum_{i=1}^n \left(\max_{j\in[k]}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j\rangle - \max_{j\in[k]}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j^\star\rangle \right)^2}_{\ell^{\mathrm{clean}}(\boldsymbol{\beta})} \\
&\quad - \underbrace{\left(\frac{1}{n}\sum_{i=1}^n z_i \left(\max_{j\in[k]}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j\rangle - \max_{j\in[k]}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j^\star\rangle \right) - \frac{1}{2n}\sum_{i=1}^n z_i^2 \right)}_{\ell^{\mathrm{noise}}(\boldsymbol{\beta})}.
\end{aligned}$$

Then the partial gradient of $\ell(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}_l$ is written as

$$\begin{aligned}\nabla_{\boldsymbol{\beta}_l}\ell(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_l\}}\left(\max_{j\in[k]}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j\rangle - \max_{j\in[k]}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j^\star\rangle - z_i\right)\boldsymbol{\xi}_i \\ &= \underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_l\}}\left(\max_{j\in[k]}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j\rangle - \max_{j\in[k]}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j^\star\rangle\right)\boldsymbol{\xi}_i}_{\nabla_{\boldsymbol{\beta}_l}\ell^{\mathrm{clean}}(\boldsymbol{\beta})} - \underbrace{\frac{1}{n}\sum_{i=1}^{n} z_i\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_l\}}\boldsymbol{\xi}_i}_{\nabla_{\boldsymbol{\beta}_l}\ell^{\mathrm{noise}}(\boldsymbol{\beta})}\end{aligned} \tag{SM3.1}$$

where $\mathcal{C}_1,\ldots,\mathcal{C}_k$ are determined by $\boldsymbol{\beta}$ as in (2.4).

In the remainder of the proof, we will use the following shorthand notation to denote the pairwise difference of parameter vectors and the probability measure on the largest partition by the ground-truth model:

$$\boldsymbol{v}_{j,j'} := \boldsymbol{\beta}_j - \boldsymbol{\beta}_{j'}, \quad \boldsymbol{v}_{j,j'}^\star := \boldsymbol{\beta}_j^\star - \boldsymbol{\beta}_{j'}^\star, \quad \text{and} \quad \pi_{\max} := \max_{j\in[k]}\mathbb{P}\left(\boldsymbol{x}\in\mathcal{C}_j^\star\right).$$

Below we show that the following lemmas hold under the condition in (2.15). The proof is provided in Appendix SM3.1.

Lemma SM3.1. *Under the hypothesis of Theorem 2.1, if (2.15) is satisfied, then with probability at least $1-\delta$ the following inequalities hold for all $j\in[k]$, $\boldsymbol{\beta}^\star\in\mathbb{R}^{k(d+1)}$, and $\boldsymbol{\beta}^t\in\mathcal{N}(\boldsymbol{\beta}^\star)$:*

$$\langle\nabla_{\boldsymbol{\beta}_j}\ell^{\mathrm{clean}}(\boldsymbol{\beta}^t),\boldsymbol{\beta}_j^t-\boldsymbol{\beta}_j^\star\rangle \geq \frac{2}{\gamma}\left(\frac{\pi_{\min}}{16}\right)^{1+\zeta^{-1}}\left(\|\boldsymbol{\beta}_j^t-\boldsymbol{\beta}_j^\star\|_2^2 - \frac{1}{10k}\sum_{j':j'\neq j}\left\|\boldsymbol{v}_{j,j'}^t-\boldsymbol{v}_{j,j'}^\star\right\|_2^2\right), \tag{SM3.2}$$

$$\|\nabla_{\boldsymbol{\beta}_j}\ell^{\mathrm{clean}}(\boldsymbol{\beta}^t)\|_2^2 \lesssim \left(\pi_{\max}+\pi_{\min}^{2(1+\zeta^{-1})}\right)\left\|\boldsymbol{\beta}_j^t-\boldsymbol{\beta}_j^\star\right\|_2^2 + \frac{\pi_{\min}^{2(1+\zeta^{-1})}}{k^2}\sum_{j':j'\neq j}\left\|\boldsymbol{v}_{j,j'}^t-\boldsymbol{v}_{j,j'}^\star\right\|_2^2, \tag{SM3.3}$$

and

$$\left\|\nabla_{\boldsymbol{\beta}_j}\ell^{\mathrm{noise}}(\boldsymbol{\beta}^t)\right\|_2 \lesssim \frac{\sigma\sqrt{kd\log(n/d)+\log(1/\delta)}}{\sqrt{n}}. \tag{SM3.4}$$

The remainder of the proof shows that the assertion of the theorem is obtained from (SM3.2), (SM3.3) and (SM3.4) via the following three steps.

Step 1: We prove by induction that all iterates remain within the neighborhood $\mathcal{N}(\boldsymbol{\beta}^\star)$. Suppose that $\boldsymbol{\beta}^t\in\mathcal{N}(\boldsymbol{\beta}^\star)$ holds for a fixed $t\in\mathbb{N}$. By the triangle inequality, for any $j\in[k]$, the next iterate $\boldsymbol{\beta}^{t+1}$ satisfies

$$\begin{aligned}\|\boldsymbol{\beta}_j^{t+1}-\boldsymbol{\beta}_j^\star\|_2 &= \|\boldsymbol{\beta}_j^t-\mu\nabla_{\boldsymbol{\beta}_j}\ell(\boldsymbol{\beta}^t)-\boldsymbol{\beta}_j^\star\|_2 \\ &\leq \underbrace{\|\boldsymbol{\beta}_j^t-\mu\nabla_{\boldsymbol{\beta}_j}\ell^{\mathrm{clean}}(\boldsymbol{\beta}^t)-\boldsymbol{\beta}_j^\star\|_2}_{A_{\mathrm{clean}}} + \underbrace{\mu\|\nabla_{\boldsymbol{\beta}_j}\ell^{\mathrm{noise}}(\boldsymbol{\beta}^t)\|_2}_{A_{\mathrm{noise}}}.\end{aligned} \tag{SM3.5}$$

Then it remains to show

$$\|\boldsymbol{\beta}_j^{t+1} - \boldsymbol{\beta}_j^\star\|_2 \le A_{\text{clean}} + A_{\text{noise}} \le \kappa\rho, \quad \forall j \in [k]. \tag{SM3.6}$$

Note that the first summand in the right-hand side of (SM3.5) satisfies

$$A_{\text{clean}}^2 = \|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2^2 - 2\mu\langle\nabla_{\boldsymbol{\beta}_j}\ell^{\text{clean}}(\boldsymbol{\beta}^t), \boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\rangle + \mu^2\|\nabla_{\boldsymbol{\beta}_j}\ell^{\text{clean}}(\boldsymbol{\beta}^t)\|_2^2.$$

Therefore, it follows from (SM3.2) and (SM3.3) that

$$\begin{aligned}
A_{\text{clean}}^2 &\le \left\|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\right\|_2^2 - \frac{4\mu}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}} \pi_{\min}^{1+\zeta^{-1}} \left(\|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2^2 - \frac{1}{10k}\sum_{j':j'\neq j}\left\|\boldsymbol{v}_{j,j'}^t - \boldsymbol{v}_{j,j'}^\star\right\|_2^2\right) \\
&\quad + \mu^2 C_1\left(\left(\pi_{\max} + \pi_{\min}^{2(1+\zeta^{-1})}\right)\left\|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\right\|_2^2 + \frac{\pi_{\min}^{2(1+\zeta^{-1})}}{k^2}\sum_{j':j'\neq j}\left\|\boldsymbol{v}_{j,j'}^t - \boldsymbol{v}_{j,j'}^\star\right\|_2^2\right) \\
&= \left(1 - \frac{4}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}} \mu\pi_{\min}^{1+\zeta^{-1}} + C_1\mu^2\left(\pi_{\max} + \pi_{\min}^{2(1+\zeta^{-1})}\right)\right)\|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2^2 \\
&\quad + \left(\frac{\frac{2}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}}\mu\pi_{\min}^{1+\zeta^{-1}}}{5k} + \frac{C_1\mu^2\pi_{\min}^{2(1+\zeta^{-1})}}{k^2}\right)\sum_{j'^\star:j'\neq j}\left\|\boldsymbol{v}_{j,j}^t - \boldsymbol{v}_{j,j'}^\star\right\|_2^2.
\end{aligned} \tag{SM3.7}$$

We set the step size μ to be

$$\mu = \frac{\omega\pi_{\min}^{1+\zeta^{-1}}}{\tau} \tag{SM3.8}$$

where ω is a constant that will be specified later and τ is given by

$$\tau := \pi_{\max} + \pi_{\min}^{2(1+\zeta^{-1})}. \tag{SM3.9}$$

Putting the choices of μ and τ respectively by (SM3.8) and (SM3.9) into (SM3.7) yields

(SM3.10)

$$\begin{aligned}
A_{\text{clean}}^2 &\le \left(1 - \frac{\frac{4}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}}\omega\pi_{\min}^{2(1+\zeta^{-1})}}{\tau} + \frac{C_1\omega^2\pi_{\min}^{2(1+\zeta^{-1})}\left(\pi_{\max} + \pi_{\min}^{2(1+\zeta^{-1})}\right)}{\tau^2}\right)\|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2^2 \\
&\quad + \left(\frac{\frac{2}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}}\omega\pi_{\min}^{2(1+\zeta^{-1})}}{5\tau k} + \frac{C_1\omega^2\pi_{\min}^{4(1+\zeta^{-1})}}{\tau^2 k^2}\right)\sum_{j':j'\neq j}\left\|\boldsymbol{v}_{j,j'}^t - \boldsymbol{v}_{j,j'}^\star\right\|_2^2 \\
&\le \left(1 - \frac{\frac{4}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}}\omega\pi_{\min}^{2(1+\zeta^{-1})}}{\tau} + \frac{C_1\omega^2\pi_{\min}^{2(1+\zeta^{-1})}}{\tau}\right)\|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2^2 \\
&\quad + \left(\frac{\frac{2}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}}\omega\pi_{\min}^{2(1+\zeta^{-1})}}{5\tau} + \frac{C_1 w^2\pi_{\min}^{2(1+\zeta^{-1})}}{\tau}\right)\max_{1\le j\neq j'\le k}\left\|\boldsymbol{v}_{j,j'}^t - \boldsymbol{v}_{j,j'}^\star\right\|_2^2.
\end{aligned}$$

Next, since $\boldsymbol{\beta}^t \in \mathcal{N}(\boldsymbol{\beta}^\star)$, by the definition of $\mathcal{N}(\boldsymbol{\beta}^\star)$ in (2.12), we have

$$\max_{j\in[k]} \|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2 \leq \kappa\rho. \tag{SM3.11}$$

Furthermore, by Lemma SM2.4, we also have

$$\max_{1\leq j\neq j'\leq k} \left\|\boldsymbol{v}_{j,j'}^t - \boldsymbol{v}_{j,j'}^\star\right\|_2 \leq 2\kappa\rho. \tag{SM3.12}$$

Then plugging in (SM3.11) and (SM3.12) into (SM3.10) yields

$$\begin{aligned}
(\kappa\rho)^{-2}A_{\text{clean}}^2 &\leq 1 - \frac{\pi_{\min}^{2(1+\zeta^{-1})}\omega}{\tau}\left(\frac{2}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}}\left(2-\frac{4}{5}\right) + C_1\omega\left(1+4\right)\right) \\
&\leq 1 - \frac{\pi_{\min}^{2(1+\zeta^{-1})}}{\tau}\cdot\omega\left(\frac{\frac{12}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}}}{5} + 5\omega C_1\right) \\
&\leq 1 - \frac{\pi_{\min}^{2(1+\zeta^{-1})}}{\tau}\cdot\omega\underbrace{\left(\frac{\frac{12}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}}}{5}\right)}_{c_0},
\end{aligned} \tag{SM3.13}$$

which is rewritten as

$$A_{\text{clean}}^2 \leq (\kappa\rho)^2\left(1 - \frac{c_0\omega\pi_{\min}^{2(1+\zeta^{-1})}}{\tau}\right). \tag{SM3.14}$$

For fixed γ and ζ, c_0 is a positive numerical constant. Due to the choice of τ by (SM3.9), we have

$$\frac{\pi_{\min}^{2(1+\zeta^{-1})}}{\tau} = \frac{\pi_{\min}^{2(1+\zeta^{-1})}}{\pi_{\max} + \pi_{\min}^{2(1+\zeta^{-1})}} < 1,$$

Furthermore, one can choose $\omega > 0$ sufficiently small so that $\omega c_0 < 1$. Then the upper bound in the right-hand side of (SM3.14) is valid as a positive number.

If A_{noise} is upper-bounded as

$$A_{\text{noise}} \leq \kappa\rho\frac{c_0\omega\pi_{\min}^{2(1+\zeta^{-1})}}{2\tau}, \tag{SM3.15}$$

then, by the elementary inequality $1-\sqrt{1-\alpha} \geq \alpha/2$ that holds for any $\alpha \in (0,1)$, we have

$$A_{\text{noise}} \leq \kappa\rho\left(1 - \sqrt{1 - \frac{c_0\omega\pi_{\min}^{2(1+\zeta^{-1})}}{\tau}}\right). \tag{SM3.16}$$

Then (SM3.14) and (SM3.16) yield (SM3.6). Therefore, it suffices to show that (SM3.15) holds.

Due to the inequality in (SM3.4), we have

$$\left\|\nabla_{\boldsymbol{\beta}_j}\ell^{\text{noise}}(\boldsymbol{\beta}^t)\right\|_2 \lesssim \frac{\sigma\sqrt{kd\log(n/d)+\log(1/\delta)}}{\sqrt{n}}, \quad \forall j\in[k].$$

By the choice of μ in (SM3.8), we obtain an upper bound on A_{noise} given by

$$A_{\text{noise}} = \mu\left\|\nabla_{\boldsymbol{\beta}_j}\ell^{\text{noise}}(\boldsymbol{\beta}^t)\right\|_2 \lesssim \frac{\omega\pi_{\min}^{1+\zeta^{-1}}}{\tau}\cdot\frac{\sigma\sqrt{kd\log(n/d)+\log(1/\delta)}}{\sqrt{n}}. \tag{SM3.17}$$

The condition in (2.15) implies

$$n \geq C\cdot\frac{\sigma^2\pi_{\min}^{-2(1+\zeta^{-1})}\left(kd\log(n/d)+\log(1/\delta)\right)}{\kappa^2\rho^2}. \tag{SM3.18}$$

One can choose the absolute constant $C>0$ in (2.15) and (SM3.18) as large enough so that (SM3.18) and (SM3.17) imply (SM3.15). This completes the induction argument in Step 1.

Step 2: Next we show that all iterates also satisfy

$$\left\|\boldsymbol{\beta}^{t+1}-\boldsymbol{\beta}^\star\right\|_2 \leq \sqrt{1-\nu}\left\|\boldsymbol{\beta}^{t}-\boldsymbol{\beta}^\star\right\|_2 + C'\mu\sigma\sqrt{\frac{k\left(kd\log(n/d)+\log(1/\delta)\right)}{n}}. \tag{SM3.19}$$

We use the fact that $\boldsymbol{\beta}^t\in\mathcal{N}(\boldsymbol{\beta}^\star)$, which has been shown in Step 1. By the update rule of gradient descent and the triangle inequality, the left-hand side of (SM3.19) satisfies

$$\begin{aligned}
\|\boldsymbol{\beta}^{t+1}-\boldsymbol{\beta}^\star\|_2 &= \|\boldsymbol{\beta}^{t}-\mu\nabla_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta}^t)-\boldsymbol{\beta}^\star\|_2\\
&\leq \|\boldsymbol{\beta}^{t}-\mu\nabla_{\boldsymbol{\beta}}\ell^{\text{clean}}(\boldsymbol{\beta}^t)-\boldsymbol{\beta}^\star\|_2 + \mu\|\nabla_{\boldsymbol{\beta}}\ell^{\text{noise}}(\boldsymbol{\beta}^t)\|_2\\
&= \underbrace{\sqrt{\sum_{j=1}^k\|\boldsymbol{\beta}_j^t-\boldsymbol{\beta}_j^\star-\mu\nabla_{\boldsymbol{\beta}_j}\ell^{\text{clean}}(\boldsymbol{\beta}^t)\|_2^2}}_{B_{\text{clean}}} + \underbrace{\sqrt{\mu^2\sum_{j=1}^k\|\nabla_{\boldsymbol{\beta}_j}\ell^{\text{noise}}(\boldsymbol{\beta}^t)\|_2^2}}_{B_{\text{noise}}}.
\end{aligned} \tag{SM3.20}$$

Below we derive an upper bound on each of the summands on the right-hand side of (SM3.20). First we show that

$$B_{\text{clean}}^2 \leq (1-\nu)\sum_{j=1}^k\left\|\boldsymbol{\beta}_j^t-\boldsymbol{\beta}_j^\star\right\|_2^2. \tag{SM3.21}$$

Since $\boldsymbol{\beta}^t\in\mathcal{N}(\boldsymbol{\beta}^\star)$, the inequality in (SM3.21) holds if there exist constants $\mu,\lambda\in(0,1)$ such that

(SM3.22)

$$\sum_{j=1}^k\langle\nabla_{\boldsymbol{\beta}_j}\ell^{\text{clean}}(\boldsymbol{\beta}^t),\boldsymbol{\beta}_j-\boldsymbol{\beta}_j^\star\rangle \geq \frac{\mu}{2}\sum_{j=1}^k\|\nabla_{\boldsymbol{\beta}_j}\ell^{\text{clean}}(\boldsymbol{\beta}^t)\|_2^2 + \frac{\lambda}{2}\sum_{j=1}^k\|\boldsymbol{\beta}_j^t-\boldsymbol{\beta}_j^\star\|_2^2, \quad \forall\boldsymbol{\beta}^t\in\mathcal{N}(\boldsymbol{\beta}^\star).$$

Indeed, the condition in (SM3.22) and $\boldsymbol{\beta}^t \in \mathcal{N}(\boldsymbol{\beta}^\star)$ imply

$$\begin{aligned}
B_{\text{clean}}^2 &= \sum_{j=1}^k \|\boldsymbol{\beta}_j^t - \mu \nabla_{\boldsymbol{\beta}_j} \ell^{\text{clean}}(\boldsymbol{\beta}^t) - \boldsymbol{\beta}_j^\star\|_2^2 \\
&= \sum_{j=1}^k \|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2^2 + \sum_{j=1}^k \mu^2 \|\nabla_{\boldsymbol{\beta}_j} \ell^{\text{clean}}(\boldsymbol{\beta}^t)\|_2^2 - 2\mu \sum_{j=1}^k \langle \boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star, \nabla_{\boldsymbol{\beta}_j} \ell^{\text{clean}}(\boldsymbol{\beta}^t)\rangle \\
&\le \sum_{j=1}^k \|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2^2 - \mu\lambda \sum_{j=1}^k \|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2^2 \\
&= (1-\mu\lambda) \sum_{j=1}^k \|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2^2. \qquad \text{(SM3.23)}
\end{aligned}$$

Next we show that (SM3.22) holds. Due to (SM3.2) and the elementary inequality $\|\boldsymbol{a}+\boldsymbol{b}\|_2^2 \le 2\|\boldsymbol{a}\|_2^2 + 2\|\boldsymbol{b}\|_2^2$, it holds for all $j \in [k]$ that

$$\begin{aligned}
&\langle \nabla_{\boldsymbol{\beta}_j} \ell^{\text{clean}}(\boldsymbol{\beta}^t), \boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\rangle \\
&\ge \frac{2}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}} \pi_{\min}^{1+\zeta^{-1}} \left(\|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2^2 - \frac{1}{5k} \sum_{j':j'\neq j} \left(\|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2^2 + \|\boldsymbol{\beta}_{j'}^t - \boldsymbol{\beta}_{j'}^\star\|_2^2 \right) \right). \qquad \text{(SM3.24)}
\end{aligned}$$

By taking the summation of (SM3.24) over $j \in [k]$, we obtain

$$\sum_{j=1}^k \langle \nabla_{\boldsymbol{\beta}_j} \ell^{\text{clean}}(\boldsymbol{\beta}^t), \boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\rangle \ge \frac{\frac{6}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}} \pi_{\min}^{1+\zeta^{-1}}}{5} \sum_{j=1}^k \|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2^2. \qquad \text{(SM3.25)}$$

Furthermore, by using (SM3.3) and the elementary inequality $\|\boldsymbol{a}+\boldsymbol{b}\|_2^2 \le 2\|\boldsymbol{a}\|_2^2 + 2\|\boldsymbol{b}\|_2^2$ again, we obtain

$$\begin{aligned}
\|\nabla_{\boldsymbol{\beta}_j} \ell^{\text{clean}}(\boldsymbol{\beta}^t)\|_2^2 &\le C_1 \left(\pi_{\max} + \pi_{\min}^{2(1+\zeta^{-1})}\right) \|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2^2 \\
&\quad + \frac{2C_1 \pi_{\min}^{2(1+\zeta^{-1})}}{k^2} \sum_{j':j'\neq j} \left(\|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2^2 + \|\boldsymbol{\beta}_{j'}^t - \boldsymbol{\beta}_{j'}^\star\|_2^2 \right). \qquad \text{(SM3.26)}
\end{aligned}$$

Summing the equation in (SM3.26) over $j \in [k]$ yields

$$\begin{aligned}
\sum_{j=1}^k \|\nabla_{\boldsymbol{\beta}_j} \ell^{\text{clean}}(\boldsymbol{\beta}^t)\|_2^2 &\le C_1 \left(\pi_{\max} + \pi_{\min}^{2(1+\zeta^{-1})} + \frac{4(k-1)\pi_{\min}^{2(1+\zeta^{-1})}}{k^2} \right) \sum_{j=1}^k \|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2^2 \\
&\le C_1 \left(\pi_{\max} + \pi_{\min}^{2(1+\zeta^{-1})} + 4\pi_{\min}^{2(1+\zeta^{-1})} \right) \sum_{j=1}^k \|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2^2. \qquad \text{(SM3.27)}
\end{aligned}$$

By combining (SM3.25) and (SM3.27) with μ as in (SM3.8), we obtain a sufficient condition for (SM3.22) given by

$$\text{(SM3.28)} \qquad \frac{\frac{6}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}}\pi_{\min}^{1+\zeta^{-1}}}{5} \geq \frac{\omega\pi_{\min}^{1+\zeta^{-1}}C_1\left(\pi_{\max}+5\pi_{\min}^{2(1+\zeta^{-1})}\right)}{2\left(\pi_{\max}+\pi_{\min}^{2(1+\zeta^{-1})}\right)}+\frac{\lambda}{2}.$$

By choosing $\omega > 0$ small enough, (SM3.28) is satisfied when λ is chosen as

$$\text{(SM3.29)} \qquad \lambda = \min(c_2\pi_{\min}^{1+\zeta^{-1}}, 1)$$

for an absolute constant $c_2 > 0$. Hence, we have shown that the condition in (SM3.22) holds with μ and λ specified by (SM3.8) and (SM3.29).

Next we consider the second summand on the right-hand side of (SM3.20). The inequality in (SM3.4) implies

$$\text{(SM3.30)} \qquad B_{\text{noise}}^2 = \mu^2\sum_{j=1}^{k}\left\|\nabla_{\boldsymbol{\beta}_j}\ell^{\text{noise}}(\boldsymbol{\beta}^t)\right\|_2^2 \lesssim \frac{\mu^2\sigma^2 k(kd\log(n/d)+\log(1/\delta))}{n}.$$

Finally, plugging in (SM3.23) and (SM3.30) into (SM3.20) provides the assertion (SM3.19). This completes the proof of Step 2.

Step 3: We finish the proof of Theorem 2.1 by applying the results in Step 1 and Step 2. Plugging in the expression of $\nu = \mu\lambda$ with μ and λ as in (SM3.8) and (SM3.29) provides

$$\begin{aligned}
\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^\star\|_2 &\leq (1-\mu\lambda)^{t/2}\|\boldsymbol{\beta}^0-\boldsymbol{\beta}^\star\|_2 + C_2\cdot\frac{\mu\sigma}{1-\sqrt{1-\mu\lambda}}\cdot\sqrt{\frac{k\left(kd\log(n/d)+\log(1/\delta)\right)}{n}}\\
&\overset{\text{(a)}}{\leq} (1-\mu\lambda)^{t/2}\|\boldsymbol{\beta}^0-\boldsymbol{\beta}^\star\|_2 + C_2\cdot\frac{2\sigma}{\lambda}\cdot\sqrt{\frac{k\left(kd\log(n/d)+\log(1/\delta)\right)}{n}}\\
&\overset{\text{(b)}}{\leq} (1-\mu\lambda)^{t/2}\|\boldsymbol{\beta}^0-\boldsymbol{\beta}^\star\|_2 + C_3\cdot\frac{\sigma}{\pi_{\max}}\cdot\sqrt{\frac{k\left(kd\log(n/d)+\log(1/\delta)\right)}{n}}\\
&\overset{\text{(c)}}{\leq} (1-\mu\lambda)^{t/2}\|\boldsymbol{\beta}^0-\boldsymbol{\beta}^\star\|_2 + C_3\cdot\sigma k\sqrt{\frac{k\left(kd\log(n/d)+\log(1/\delta)\right)}{n}},
\end{aligned}$$

where (a) follows from the elementary inequality $\sqrt{1-t} < 1-t/2$ for any $t \in (0,1)$; (b) holds by the choice of τ in (SM3.9); (c) holds since $\pi_{\max}^{-1} \leq k$.

SM3.1. Proof of Lemma SM3.1.

We show that each of (SM3.2), (SM3.3), and (SM3.4) holds with probability at least $1-\delta/3$. We also note that for simplicity, we proceed on the proofs using $\boldsymbol{\beta}$ and $\boldsymbol{v}_{j,j'}$. Therefore, the assertions in (SM3.2), (SM3.3), and (SM3.4) can be completed by substituting $\boldsymbol{\beta}$ and $\boldsymbol{v}_{j,j'}$ with $\boldsymbol{\beta}^t$ and $\boldsymbol{v}_{j,j'}^t$ respectively.

Proof of (SM3.2): We show that (SM3.2) holds with high probability under the following condition

$$\text{(SM3.31)} \qquad n \geq C_1\left(\log(k/\delta)\vee d\log(n/d)\right)k^4\pi_{\min}^{-4(1+\zeta^{-1})},$$

which is implied by the assumption in (2.15). We proceed with the proof under the following three events, each of which holds with probability at least $1-\delta/9$. First, since (SM3.31) implies (SM2.13), by Lemma SM2.7, it holds with probability at least $1-\delta/9$ that

$$\begin{aligned}&\frac{1}{n}\sum_{j':j'\neq j}\sum_{i=1}^{n}\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\cap\mathcal{C}^\star_{j'}\}}\langle\boldsymbol{\xi}_i,\boldsymbol{v}^\star_{j,j'}\rangle^2\\ &\leq\frac{2}{5\gamma k}\left(\frac{\pi_{\min}}{16}\right)^{1+\zeta^{-1}}\sum_{j':j'\neq j}\|\boldsymbol{v}_{j,j'}-\boldsymbol{v}^\star_{j,j'}\|_2^2,\quad\forall j\in[k],\ \forall\boldsymbol{\beta}\in\mathcal{N}(\boldsymbol{\beta}^\star),\ \forall\boldsymbol{\beta}^\star\in\mathbb{R}^{d+1}.\end{aligned}\tag{SM3.32}$$

Moreover, since (SM3.31) also implies (SM2.7), by Lemma SM2.6, it holds with probability at least $1-\delta/3$ that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\cap\mathcal{C}^\star_j\}}\geq\frac{\pi_{\min}}{4},\quad\forall j\in[k],\ \forall\boldsymbol{\beta}\in\mathcal{N}(\boldsymbol{\beta}^\star),\ \forall\boldsymbol{\beta}^\star\in\mathbb{R}^{d+1}.\tag{SM3.33}$$

Lastly, since (SM3.31) is a sufficient condition to invoke Lemma SM2.3 with $\alpha=\pi_{\min}/4$, it holds with probability at least $1-\delta/9$ that

$$\inf_{\mathcal{I}\subset[n]:|\mathcal{I}|\geq\frac{\pi_{\min}n}{4}}\lambda_{d+1}\left(\frac{1}{n}\sum_{i\in\mathcal{I}}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top\right)\geq\frac{2}{\gamma}\left(\frac{\pi_{\min}}{16}\right)^{1+\zeta^{-1}}.\tag{SM3.34}$$

Therefore, we have shown that (SM3.32), (SM3.33), and (SM3.34) hold with probability at least $1-\delta/3$. The remainder of the proof is conditioned on the event that $\{\boldsymbol{\xi}_i\}_{i=1}^n$ satisfy (SM3.32), (SM3.33), and (SM3.34).

Let $\boldsymbol{\beta}^\star\in\mathbb{R}^{d+1}$, $\boldsymbol{\beta}\in\mathcal{N}(\boldsymbol{\beta}^\star)$, and $j\in[k]$ be arbitrarily fixed. For brevity, we will use the shorthand notation $\boldsymbol{h}_j:=\boldsymbol{\beta}_j-\boldsymbol{\beta}^\star_j$. Then the left-hand side of (SM3.2) is rewritten as

$$\begin{aligned}\langle\nabla_{\boldsymbol{\beta}_j}\ell^{\text{clean}}(\boldsymbol{\beta}),\boldsymbol{h}_j\rangle&=\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\}}\left(\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j\rangle-\max_{j\in[k]}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}^\star_j\rangle\right)\langle\boldsymbol{\xi}_i,\boldsymbol{h}_j\rangle\\ &=\frac{1}{n}\sum_{j'=1}^{k}\sum_{i=1}^{n}\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\cap\mathcal{C}^\star_j\}}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j-\boldsymbol{\beta}^\star_{j'}\rangle\langle\boldsymbol{\xi}_i,\boldsymbol{h}_j\rangle\\ &=\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\cap\mathcal{C}^\star_j\}}\langle\boldsymbol{\xi}_i,\boldsymbol{h}_j\rangle^2+\frac{1}{n}\sum_{j':j'\neq j}\sum_{i=1}^{n}\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\cap\mathcal{C}^\star_{j'}\}}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j-\boldsymbol{\beta}^\star_{j'}\rangle\langle\boldsymbol{\xi}_i,\boldsymbol{h}_j\rangle.\end{aligned}$$

By the inequality of arithmetic and geometric means, we have

$$\begin{aligned}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j-\boldsymbol{\beta}^\star_{j'}\rangle\langle\boldsymbol{\xi}_i,\boldsymbol{h}_j\rangle&=\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j-\boldsymbol{\beta}^\star_j+\boldsymbol{\beta}^\star_j-\boldsymbol{\beta}^\star_{j'}\rangle\langle\boldsymbol{\xi}_i,\boldsymbol{h}_j\rangle\\ &=\langle\boldsymbol{\xi}_i,\boldsymbol{h}_j+\boldsymbol{v}^\star_{j,j'}\rangle\langle\boldsymbol{\xi}_i,\boldsymbol{h}_j\rangle\\ &\geq\frac{\langle\boldsymbol{\xi}_i,\boldsymbol{h}_j\rangle^2}{2}-\frac{\langle\boldsymbol{\xi}_i,\boldsymbol{v}^\star_{j,j'}\rangle^2}{2}\geq-\frac{\langle\boldsymbol{\xi}_i,\boldsymbol{v}^\star_{j,j'}\rangle^2}{2}.\end{aligned}$$

Therefore, we obtain

$$\langle \nabla_{\boldsymbol{\beta}_j} \ell^{\text{clean}}(\boldsymbol{\beta}), \boldsymbol{h}_j \rangle \geq \underbrace{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j \cap \mathcal{C}_j^\star\}} \langle \boldsymbol{\xi}_i, \boldsymbol{h}_j \rangle^2}_{(*)} - \underbrace{\frac{1}{2n} \sum_{j': j' \neq j} \sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j \cap \mathcal{C}_{j'}^\star\}} \langle \boldsymbol{\xi}_i, \boldsymbol{v}_{j,j'}^\star \rangle^2}_{(**)} . \tag{SM3.35}$$

By (SM3.33) and (SM3.34), the first summand in the right-hand side of (SM3.35) is bounded from below as

$$(*) \geq \frac{2}{\gamma} \left(\frac{\pi_{\min}}{16} \right)^{1+\zeta^{-1}} \|\boldsymbol{h}_j\|_2^2. \tag{SM3.36}$$

Moreover, due to (SM3.32), $(**)$ is bounded from above as

$$(**) \leq \frac{1}{5\gamma k} \left(\frac{\pi_{\min}}{16} \right)^{1+\zeta^{-1}} \sum_{j': j' \neq j} \|\boldsymbol{v}_{j,j'} - \boldsymbol{v}_{j,j'}^\star\|_2^2. \tag{SM3.37}$$

Then, plugging in (SM3.36) and (SM3.37) into (SM3.35) provides

$$\begin{aligned}
&\langle \nabla_{\boldsymbol{\beta}_j} \ell(\boldsymbol{\beta}), \boldsymbol{h}_j \rangle \\
&\geq \frac{2}{\gamma} \left(\frac{\pi_{\min}}{16} \right)^{1+\zeta^{-1}} \|\boldsymbol{h}_j\|_2^2 - \frac{1}{5\gamma} \left(\frac{1}{16} \right)^{1+\zeta^{-1}} \left(\frac{\pi_{\min}^{1+\zeta^{-1}}}{k} \right) \sum_{j': j' \neq j} \|\boldsymbol{v}_{j,j'} - \boldsymbol{v}_{j,j'}^\star\|_2^2 \\
&= \frac{2}{\gamma} \left(\frac{\pi_{\min}}{16} \right)^{1+\zeta^{-1}} \left(\|\boldsymbol{h}_j\|_2^2 - \frac{1}{10k} \sum_{j': j' \neq j} \|\boldsymbol{v}_{j,j'} - \boldsymbol{v}_{j,j'}^\star\|_2^2 \right).
\end{aligned}$$

This completes the proof.

Proof of (SM3.3): The proof is based on the condition

$$n \geq C_2 \left(\log(k/\delta) \vee d \log(n/d) \right) k^4 \pi_{\min}^{-4(1+\zeta^{-1})}, \tag{SM3.38}$$

which is implied by (2.15). We will proceed under the following four events, each of which holds with probability at least $1 - \delta/12$. First, since (SM3.38) implies (SM2.13), by Lemma SM2.7, (SM3.32) holds with probability at least $1 - \delta/12$. Next, since $\left(\mathcal{C}_j^\star \right)_{j=1}^{k}$ are included in the set of intersection of k half-spaces in $\mathbb{R}^d$, by Corollary SM1.7 and (SM3.38), it holds with probability at least $1 - \delta/12$ that

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j^\star\}} \leq 2\mathbb{P}\left(\boldsymbol{x} \in \mathcal{C}_j^\star \right), \quad \forall j \in [k]. \tag{SM3.39}$$

We also consider the event given by

$$\sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j \cap \mathcal{C}_j^\star\}} \leq 2nc \left(\frac{\pi_{\min}^{2(1+\zeta^{-1})}}{k^2} \right), \quad \forall j \neq j',\ \forall \boldsymbol{\beta} \in \mathcal{N}(\boldsymbol{\beta}^\star) \tag{SM3.40}$$

for some numerical constant $c \in (0,1)$. Note that (SM3.38) is a sufficient condition to invoke Lemma SM2.7 with probability at least $1-\delta/12$. Therefore, all intermediate steps in the proof of Lemma SM2.7 hold. In particular, due to the inclusion argument in (SM2.15), $\boldsymbol{x}_i \in \mathcal{C}_j \cap \mathcal{C}^\star_{j'}$ implies $\boldsymbol{\xi}_i = [\boldsymbol{x}_i; 1] \in \mathcal{S}_{\boldsymbol{v}_{j,j'}, \boldsymbol{v}^\star_{j,j'}}$ for any $j \neq j'$, where $\mathcal{S}_{\boldsymbol{v}_{j,j'}, \boldsymbol{v}^\star_{j,j'}}$ is defined in (SM2.17). Then, (SM2.21) with α as in (SM2.20) implies (SM3.40). The last event is defined by

(SM3.41)
$$\max_{\substack{\mathcal{I} \subset [n] \\ |\mathcal{I}| \le 2\alpha n}} \lambda_{\max}\left(\frac{1}{n}\sum_{i\in\mathcal{I}} \boldsymbol{\xi}_i\boldsymbol{\xi}_i^\intercal\right) \le C_4(\eta^2 \vee 1)\sqrt{\alpha}, \quad \forall \alpha \in \left\{\frac{c\pi_{\min}^{2(1+\zeta^{-1})}}{k^2}\right\} \cup \left\{\mathbb{P}(\boldsymbol{x} \in \mathcal{C}^\star_j)\right\}_{j=1}^k.$$

By (SM3.38), Lemma SM2.1, and the union bound over $j \in [k]$, (SM3.41) holds with probability at least $1-\delta/12$. Thus far we have shown that (SM3.32), (SM3.39), (SM3.40), and (SM3.41) hold with probability at least $1-\delta/3$. We proceed conditioned on the event that $\{\boldsymbol{\xi}_i\}_{i=1}^n$ satisfy these conditions.

Let $\boldsymbol{\beta}^\star \in \mathbb{R}^{d+1}$, $\boldsymbol{\beta} \in \mathcal{N}(\boldsymbol{\beta}^\star)$, and $j \in [k]$ be arbitrarily fixed. Then the partial gradient of $\ell^{\text{clean}}(\boldsymbol{\beta})$ with respect to the jth block $\boldsymbol{\beta}_j \in \mathbb{R}^{d+1}$ of $\boldsymbol{\beta} \in \mathbb{R}^{k(d+1)}$ is written as

$$\begin{aligned}
\nabla_{\boldsymbol{\beta}_j}\ell^{\text{clean}}(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\}}\left(\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j\rangle - \max_{j\in[k]}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}^\star_j\rangle\right)\boldsymbol{\xi}_i \\
&= \frac{1}{n}\sum_{j'\in[k]}\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\cap\mathcal{C}^\star_{j'}\}}\left(\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j\rangle - \langle\boldsymbol{\xi}_i,\boldsymbol{\beta}^\star_{j'}\rangle\right)\boldsymbol{\xi}_i \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\cap\mathcal{C}^\star_{j'}\}}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j-\boldsymbol{\beta}^\star_j\rangle\boldsymbol{\xi}_i + \frac{1}{n}\sum_{j':j'\neq j}\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\cap\mathcal{C}^\star_{j'}\}}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j-\boldsymbol{\beta}^\star_{j'}\rangle\boldsymbol{\xi}_i. \qquad \text{(SM3.42)}
\end{aligned}$$

By using the identity $\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j-\boldsymbol{\beta}^\star_{j'}\rangle = \langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j-\boldsymbol{\beta}^\star_j+\boldsymbol{\beta}^\star_j-\boldsymbol{\beta}^\star_{j'}\rangle$, (SM3.42) is rewritten as

(SM3.43)
$$\nabla_{\boldsymbol{\beta}_j}\ell^{\text{clean}}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\}}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j-\boldsymbol{\beta}^\star_j\rangle\boldsymbol{\xi}_i + \frac{1}{n}\sum_{j':j'\neq j}\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\cap\mathcal{C}^\star_{j'}\}}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}^\star_j-\boldsymbol{\beta}^\star_{j'}\rangle\boldsymbol{\xi}_i.$$

Then it follows from (SM3.43) that

$$
\begin{aligned}
&\left\|\nabla_{\boldsymbol{\beta}_j}\ell^{\mathrm{clean}}(\boldsymbol{\beta})\right\|_2^2 \\
&\overset{\text{(i)}}{\leq} 2\left\|\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\}}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j-\boldsymbol{\beta}_j^\star\rangle\boldsymbol{\xi}_i\right\|_2^2 + 2\left\|\frac{1}{n}\sum_{j':j'\neq j}\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\cap\mathcal{C}_{j'}^\star\}}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j^\star-\boldsymbol{\beta}_{j'}^\star\rangle\boldsymbol{\xi}_i\right\|_2^2 \\
&\overset{\text{(ii)}}{\leq} 2\cdot\left\|\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\}}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top\right\|\cdot\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\}}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j-\boldsymbol{\beta}_j^\star\rangle^2 \\
&\quad + 2\cdot\sum_{j':j'\neq j}\left\|\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\cap\mathcal{C}\}}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top\right\|\cdot\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\cap\mathcal{C}_{j'}^\star\}}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j^\star-\boldsymbol{\beta}_{j'}^\star\rangle^2 \\
&\leq 2\cdot\underbrace{\left\|\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\}}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top\right\|^2}_{\text{(a)}}\cdot\|\boldsymbol{\beta}_j-\boldsymbol{\beta}_j^\star\|_2^2 \\
&\quad + 2\cdot\underbrace{\max_{j':j'\neq j}\left\|\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i\mathcal{C}_j\cap\mathcal{C}_{j'}^\star\}}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top\right\|}_{\text{(b)}}\cdot\underbrace{\frac{1}{n}\sum_{j':j'\neq j}\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\cap\mathcal{C}_{j'}^\star\}}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j^\star-\boldsymbol{\beta}_{j'}^\star\rangle^2}_{\text{(c)}},
\end{aligned}
\tag{SM3.44}
$$

where (i) holds since $\|\boldsymbol{a}+\boldsymbol{b}\|_2^2 \leq 2\|\boldsymbol{a}\|_2^2 + 2\|\boldsymbol{b}\|_2^2$ and (ii) holds since $\mathcal{C}_j\cap\mathcal{C}_l^\star$ and $\mathcal{C}_j\cap\mathcal{C}_{l'}^\star$ are disjoint for any $l\neq l'\in[k]$. An upper bound on (b) is provided by (SM3.32). It remains to derive upper bounds on (a) and (c).

First, we derive an upper bound on (a). By the triangle inequality, we have

$$
\sqrt{\text{(a)}} \leq \sum_{j'=1}^k \left\|\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\cap\mathcal{C}_{j'}^\star\}}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\mathsf{T}\right\|. \tag{SM3.45}
$$

For the summand indexed by $j'=j$, due to the set inclusion $\mathcal{C}_j\cap\mathcal{C}_j^\star\subset\mathcal{C}_j^\star$, we obtain that

$$
\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\cap\mathcal{C}_j^\star\}}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\mathsf{T} \preceq \sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j^\star\}}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\mathsf{T}.
$$

Therefore, by (SM3.39) and (SM3.41), we have

$$
\begin{aligned}
\left\|\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j^\star\}}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\mathsf{T}\right\| &\leq \max_{\mathcal{I}:|\mathcal{I}|\leq 2n\mathbb{P}(\boldsymbol{x}\in\mathcal{C}_j^\star)}\left\|\frac{1}{n}\sum_{i\in\mathcal{I}}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\mathsf{T}\right\| \\
&\lesssim (\eta^2\vee 1)\sqrt{\mathbb{P}(\boldsymbol{x}\in\mathcal{C}_j^\star)} \\
&\leq (\eta^2\vee 1)\sqrt{\pi_{\max}},
\end{aligned}
\tag{SM3.46}
$$

where the last inequality holds by the definition of $\pi_{\max}$. Similarly, by (SM3.40) and (SM3.41), we have

$$\left\| \sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j \cap \mathcal{C}_{j'}^\star\}} \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\intercal \right\| \lesssim (\eta^2 \vee 1)\sqrt{c} \left(\frac{\pi_{\min}^{1+\zeta^{-1}}}{k} \right), \quad \forall j' \neq j. \tag{SM3.47}$$

Then by plugging in (SM3.46) and (SM3.47) to (SM3.45), we obtain

$$(\mathrm{a}) \lesssim \left(\pi_{\max} + \pi_{\min}^{2(1+\zeta^{-1})} \right) \left\| \boldsymbol{\beta}_j - \boldsymbol{\beta}_j^\star \right\|_2^2$$

for an absolute constant C_1. Finally, since an upper bound on (b) is given by (SM3.47), plugging in the obtained upper bounds to (SM3.44) provides the assertion.

Proof of (SM3.4): By the variational characterization of the Euclidean norm and the triangle inequality, we have

$$\begin{aligned} \left\| \nabla_{\boldsymbol{\beta}_j} \ell^{\mathrm{noise}}(\boldsymbol{\beta}) \right\|_2 &= \sup_{[\boldsymbol{u};\ w] \in B_2^{d+1}} \left| \frac{1}{n} \sum_{i=1}^{n} z_i \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j\}} (\langle \boldsymbol{x}_i, \boldsymbol{u} \rangle + w) \right| \\ &\leq \underbrace{\sup_{\boldsymbol{u} \in B_2^p} \left| \frac{1}{n} \sum_{i=1}^{n} z_i \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j\}} \langle \boldsymbol{x}_i, \boldsymbol{u} \rangle \right|}_{(\mathrm{A})} + \underbrace{\sup_{|w| \leq 1} \left| \frac{1}{n} \sum_{i=1}^{n} z_i \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j\}} w \right|}_{(\mathrm{B})}, \end{aligned} \tag{SM3.48}$$

where B_2^d denotes the unit ball in ℓ_2^d. Note that (A) and (B) depend on $\boldsymbol{\beta}$ only through $\mathcal{C}_j$, which are determined by $\boldsymbol{\beta}$ according to (2.4). For any $\boldsymbol{\beta}$ and any $j \in [k]$, the corresponding $\mathcal{C}_j$ is given as the intersection of up to k affine spaces. Therefore, it suffices to maximize $\left\| \nabla_{\boldsymbol{\beta}_j} \ell^{\mathrm{noise}}(\boldsymbol{\beta}) \right\|_2$ over $\mathcal{C}_j \in \mathcal{P}_{k-1}$ for a fixed j, where $\mathcal{P}_{k-1}$ is defined in the statement of Lemma SM1.6.

We proceed under the event that the following inequalities hold:

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\intercal \right\| \leq 1 + \epsilon \tag{SM3.49}$$

and

$$\left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j\}} - \mathbb{P}(\boldsymbol{x} \in \mathcal{C}_j) \right| \leq \epsilon, \quad \forall \mathcal{C}_j \in \mathcal{P}_{k-1} \tag{SM3.50}$$

for some constant ϵ, which we specify later. The remainder of the proof is given conditioned on $(\boldsymbol{x}_i)_{i=1}^n$ satisfying (SM3.49) and (SM3.50).

First, we derive an upper bound on (A) in (SM3.48). Note that (A) corresponds to the supremum of the random process

$$Z_{\boldsymbol{u}} := \frac{1}{n} \sum_{i=1}^{n} z_i \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j\}} \langle \boldsymbol{x}_i, \boldsymbol{u} \rangle$$

over $\boldsymbol{u} \in B_2^p$. The sub-Gaussian increment satisfies

$$\begin{aligned}\|Z_{\boldsymbol{u}} - Z_{\boldsymbol{u}'}\|_{\psi_2} &\lesssim \frac{\sigma}{\sqrt{n}} \sqrt{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j\}} \langle \boldsymbol{x}_i, \boldsymbol{u} - \boldsymbol{u}' \rangle^2} \\ &\le \frac{\sigma}{\sqrt{n}} \left\| \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j\}} \boldsymbol{x}_i \boldsymbol{x}_i^\mathsf{T} \right\|^{1/2} \cdot \|\boldsymbol{u} - \boldsymbol{u}'\|_2 \\ &\le \frac{\sigma}{\sqrt{n}} \left\| \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\mathsf{T} \right\|^{1/2} \cdot \|\boldsymbol{u} - \boldsymbol{u}'\|_2 \\ &\le \frac{\sigma\sqrt{1+\epsilon}}{\sqrt{n}} \cdot \|\boldsymbol{u} - \boldsymbol{u}'\|_2,\end{aligned}$$

where the third step follows from the inequality

$$\left\| \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j\}} \boldsymbol{x}_i \boldsymbol{x}_i^\mathsf{T} \right\| \le \left\| \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\mathsf{T} \right\|,$$

which holds deterministically, and the last step follows from (SM3.49). Then, by applying a version of Dudley's inequality [SM11, Theorem 8.1.6], we obtain that

$$\mathbb{P}\left(\sup_{\boldsymbol{u} \in B_2^p} |Z_{\boldsymbol{u}}| > \frac{C_1 \sigma\sqrt{1+\epsilon}}{\sqrt{n}} \left(\int_0^\infty \sqrt{\log N(B_2^p, \|\cdot\|_2, \eta)}\, d\eta + \sqrt{\log(1/\delta)} \right) \right) \le \delta.$$

By the elementary upper bound on the covering number $N(B_2^p, \|\cdot\|_2, \eta) \le (3/\eta)^p$ (e.g. see [SM11, Example 8.1.11]) and the definition of (A) in (SM3.48), we have

$$\text{(A)} \lesssim \sqrt{\frac{\sigma^2(1+\epsilon)(d + \log(1/\delta))}{n}}, \tag{SM3.51}$$

holds with probability $1 - \delta/3$. Then we apply the union bound over $\mathcal{C}_j \in \mathcal{P}_{k-1}$. It follows from (SM1.1) that

$$\sup_{\mathcal{C}_j \in \mathcal{P}_{k-1}} \text{(A)} \lesssim \sqrt{\frac{\sigma^2(1+\epsilon)(\log(1/\delta) + kd\log(n/d))}{n}}$$

holds with probability $1 - \delta/9$.

Next we derive an upper bound on (B) in (SM3.48). Note that (B) is rewritten as the absolute value of

$$\varrho = \frac{1}{n}\sum_{i=1}^n z_i \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j\}}.$$

Conditioned on $(\boldsymbol{x}_i)_{i=1}^n$ satisfying (SM3.50), ϱ is a sub-Gaussian random variable that satisfies $\mathbb{E}\varrho = 0$ and

$$\mathbb{E}\varrho^2 = \frac{\sigma^2}{n} \cdot \left(\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j\}} \right) \le \frac{\sigma^2(\mathbb{P}(\boldsymbol{x} \in \mathcal{C}_j) + \epsilon)}{n}.$$

The standard sub-Gaussian tail bound implies

$$\mathbb{P}\left(|\varrho| > \sqrt{\frac{C_2\sigma^2(\mathbb{P}(\boldsymbol{x}\in\mathcal{C}_j)+\epsilon)\log(1/\delta)}{n}}\right) \le \delta.$$

By taking the union bound over $\mathcal{C}_j \in \mathcal{P}_{k-1}$ and utilizing the inequality in (SM1.1), we obtain that

$$\sup_{\mathcal{C}_j\in\mathcal{P}_{k-1}} (\mathrm{B}) \lesssim \sqrt{\frac{\sigma^2(\mathbb{P}(\boldsymbol{x}\in\mathcal{C}_j)+\epsilon)\,(kd\log(n/d)+\log(1/\delta))}{n}}$$

$$\le \sqrt{\frac{\sigma^2(1+\epsilon)\,(kd\log(n/d)+\log(1/\delta))}{n}} \tag{SM3.52}$$

holds with probability $1-\delta/9$.

Finally it remains to show that (SM3.49) and (SM3.50) hold with probability $1-\delta/3$ for ϵ satisfying

$$\epsilon \lesssim \sqrt{\frac{kp(\log(n/d)+\log(1/\delta))}{n}}.$$

This is obtained as a direct consequence of Lemmas SM1.1 and SM1.3. One can choose the absolute constant C in (2.15) large enough so that $\epsilon < 1$. Then the parameter ϵ in (SM3.51) and (SM3.52) will be dropped. This completes the proof.

SM4. Proof of Theorem 3.1.

The proof will be similar to that for Theorem 2.1. We will focus on the distinction due to the modification of the algorithm with random sampling. The partial subgradient in the update for the mini-batch stochastic gradient descent algorithm is given by

$$\frac{1}{m}\sum_{i\in I_t}\nabla_{\boldsymbol{\beta}_l}\ell_i(\boldsymbol{\beta}^t) = \frac{1}{m}\sum_{i\in I_t}\underbrace{\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_l\}}\left(\max_{j\in[k]}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j^t\rangle - \max_{j\in[k]}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j^\star\rangle\right)\boldsymbol{\xi}_i}_{\nabla_{\boldsymbol{\beta}_l}\ell_i^{\mathrm{clean}}(\boldsymbol{\beta}^t)} - \frac{1}{m}\sum_{i\in I_t}\underbrace{z_i\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_l\}}\boldsymbol{\xi}_i}_{\nabla_{\boldsymbol{\beta}_l}\ell_i^{\mathrm{noise}}(\boldsymbol{\beta}^t)},$$

where $\mathcal{C}_1,\dots,\mathcal{C}_k$ are determined by $\boldsymbol{\beta}^t$ as in (2.4).

As shown in Section SM3, (2.15) invokes Lemma SM3.1 and hence (SM3.2) holds with probability $1-\delta/3$. Next, we show that under the condition (2.15), the statements of the following lemma hold with probability $1-2\delta/3$. The proof is provided in Appendix SM4.1.

Lemma SM4.1. *Suppose that the hypothesis of Theorem 3.1 holds. If (2.15) is satisfied, then the following statement holds with probability at least $1-2\delta/3$: For all $j\in[k]$, $\boldsymbol{\beta}^\star\in\mathbb{R}^{k(d+1)}$, and $\boldsymbol{\beta}^t\in\mathcal{N}(\boldsymbol{\beta}^\star)$, we have*

$$\mathbb{E}_{I_t}\left\|\frac{1}{m}\sum_{i\in I_t}\nabla_{\boldsymbol{\beta}_j}\ell_i^{\mathrm{clean}}(\boldsymbol{\beta}^t)\right\|_2^2 \lesssim \left(1\vee\frac{d+\log(n/\delta)}{m}\right)\left(\left(\sqrt{\pi_{\max}}+\pi_{\min}^{1+\zeta^{-1}}\right)\left\|\boldsymbol{\beta}_j^t-\boldsymbol{\beta}_j^\star\right\|_2^2 + \frac{\pi_{\min}^{1+\zeta^{-1}}}{k}\sum_{j':j'\neq j}\left\|\boldsymbol{v}_{j,j'}^t-\boldsymbol{v}_{j,j'}^\star\right\|_2^2\right), \tag{SM4.1}$$

and

$$\mathbb{E}_{I_t}\left\|\frac{1}{m}\sum_{i\in I_t}\nabla_{\boldsymbol{\beta}_j}\ell_i^{\text{noise}}(\boldsymbol{\beta}^t)\right\|_2^2 \lesssim \sigma^2\left(\frac{d+\log(n/\delta)}{m}\vee\frac{kd\log(n/d)+\log(1/\delta)}{n}\right). \tag{SM4.2}$$

Then we show that the assertion of the theorem follows from (SM3.2), (SM4.1), and (SM4.2) via the following three steps.

Step 1: We show that every iterate remains within the neighborhood $\mathcal{N}(\boldsymbol{\beta}^\star)$ by the induction argument. Therefore, we illustrate that if we suppose $\boldsymbol{\beta}^t \in \mathcal{N}(\boldsymbol{\beta}^\star)$ holds for a fixed $t \in \mathbb{N}$, we show $\boldsymbol{\beta}^{t+1} \in \mathcal{N}(\boldsymbol{\beta}^\star)$ in expectation. By the update rule of SGD with batch size m, the triangle inequality gives

$$\mathbb{E}_{I_t}\|\boldsymbol{\beta}_j^{t+1}-\boldsymbol{\beta}_j^\star\|_2 \le \underbrace{\mathbb{E}_{I_t}\left\|\boldsymbol{\beta}_j^t-\mu\frac{1}{m}\sum_{i\in I_t}\nabla_{\boldsymbol{\beta}_j}\ell_i^{\text{clean}}(\boldsymbol{\beta}^t)-\boldsymbol{\beta}_j^\star\right\|_2}_{A_{\text{clean}}} + \underbrace{\mu\mathbb{E}_{I_t}\left\|\frac{1}{m}\sum_{i\in I_t}\nabla_{\boldsymbol{\beta}_j}\ell_i^{\text{noise}}(\boldsymbol{\beta}^t)\right\|_2}_{A_{\text{noise}}}. \tag{SM4.3}$$

We will show that

$$\mathbb{E}_{I_t}\|\boldsymbol{\beta}_j^{t+1}-\boldsymbol{\beta}_j^\star\|_2 \le A_{\text{clean}}+A_{\text{noise}} \le \kappa\rho, \quad \forall j\in[k]. \tag{SM4.4}$$

By applying Jensen's inequality, we can obtain an upper-bound A_{clean} in (SM4.3):

$$\begin{aligned} A_{\text{clean}}^2 &\le \mathbb{E}_{I_t}\left\|\boldsymbol{\beta}_j^t-\mu\cdot\frac{1}{m}\sum_{i\in I_t}\nabla_{\boldsymbol{\beta}_j}\ell_i^{\text{clean}}(\boldsymbol{\beta}^t)-\boldsymbol{\beta}_j^\star\right\|_2^2 \\ &= \|\boldsymbol{\beta}_j^t-\boldsymbol{\beta}_j^\star\|_2^2 - 2\mu\mathbb{E}_{I_t}\left\langle\frac{1}{m}\sum_{i\in I_t}\nabla_{\boldsymbol{\beta}_j}\ell_i^{\text{clean}}(\boldsymbol{\beta}^t),\boldsymbol{\beta}_j^t-\boldsymbol{\beta}_j^\star\right\rangle + \mu^2\mathbb{E}_{I_t}\left\|\frac{1}{m}\sum_{i\in I_t}\nabla_{\boldsymbol{\beta}_j}\ell_i(\boldsymbol{\beta}^t)\right\|_2^2. \end{aligned} \tag{SM4.5}$$

Due to the expectation, the second term in (SM4.5) simplifies to

$$\mathbb{E}_{I_t}\left\langle\frac{1}{m}\sum_{i\in I_t}\nabla_{\boldsymbol{\beta}_j}\ell_i^{\text{clean}}(\boldsymbol{\beta}^t),\boldsymbol{\beta}_j^t-\boldsymbol{\beta}_j^\star\right\rangle = \langle\nabla_{\boldsymbol{\beta}_j}\ell^{\text{clean}}(\boldsymbol{\beta}^t),\boldsymbol{\beta}_j^t-\boldsymbol{\beta}_j^\star\rangle, \tag{SM4.6}$$

where $\nabla_{\boldsymbol{\beta}_j}\ell^{\text{clean}}(\boldsymbol{\beta}^t)$ is defined in (SM3.1). Then, (SM3.2) gives a lower bound on (SM4.6). Furthermore, an upper bound on the third term in (SM4.5) is given by (SM4.1). Putting the bounds (SM3.2) and (SM4.1) in (SM4.5) provides

$$
\begin{aligned}
&A_{\text{clean}}^2 \leq \\
&\left(1 - \frac{4}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}} \mu\pi_{\min}^{1+\zeta^{-1}} + C_1\mu^2\left(1 \vee \frac{d+\log(n/\delta)}{m}\right)\left(\sqrt{\pi_{\max}} + \pi_{\min}^{1+\zeta^{-1}}\right)\right)\|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2^2 \\
&\quad + \left(\frac{\frac{2}{\gamma}\left(\frac{1}{16}^{1+\zeta^{-1}}\right)\mu\pi_{\min}^{1+\zeta^{-1}}}{5k} + C_1\left(1 \vee \frac{d+\log(n/\delta)}{m}\right)\frac{\mu^2\pi_{\min}^{1+\zeta^{-1}}}{k}\right)\sum_{j'^*:j'\neq j}\left\|\boldsymbol{v}_{j,j}^t - \boldsymbol{v}_{j,j'}^\star\right\|_2^2.
\end{aligned}
\tag{SM4.7}
$$

Let us choose the step size μ following

$$
\mu = \frac{\omega\pi_{\min}^{1+\zeta^{-1}}}{\tau}\cdot\left(1 \wedge \frac{m}{d+\log(n/\delta)}\right)
\tag{SM4.8}
$$

for a numerical constant ω, which we specify later, and τ defined as

$$
\tau := \sqrt{\pi_{\max}} + \pi_{\min}^{1+\zeta^{-1}}.
\tag{SM4.9}
$$

Taking μ by (SM4.8) and τ by (SM4.9) in (SM4.7) yields

$$
\begin{aligned}
&A_{\text{clean}}^2 \\
&\leq \left(1 - \left(1 \wedge \frac{m}{d+\log(n/\delta)}\right)\cdot\right. \\
&\qquad\left.\left(\frac{\frac{4}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}}\omega\pi_{\min}^{2(1+\zeta^{-1})}}{\tau} - \frac{C_1\omega^2\pi_{\min}^{2(1+\zeta^{-1})}\left(\sqrt{\pi_{\max}} + \pi_{\min}^{1+\zeta^{-1}}\right)}{\tau^2}\right)\right)\|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2^2 \\
&\quad + \left(1 \wedge \frac{m}{d+\log(n/\delta)}\right)\cdot\left(\frac{\frac{2}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}}\omega\pi_{\min}^{2(1+\zeta^{-1})}}{5\tau k} + \frac{C_1\omega^2\pi_{\min}^{3(1+\zeta^{-1})}}{\tau^2 k}\right)\sum_{j':j'\neq j}\left\|\boldsymbol{v}_{j,j'}^t - \boldsymbol{v}_{j,j'}^\star\right\|_2^2 \\
&\leq \left(1 - \left(1 \wedge \frac{m}{d+\log(n/\delta)}\right)\cdot\left(\frac{\frac{4}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}}\omega\pi_{\min}^{2(1+\zeta^{-1})}}{\tau} - \frac{C_1\omega^2\pi_{\min}^{2(1+\zeta^{-1})}}{\tau}\right)\right)\|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2^2 \\
&\quad + \left(1 \wedge \frac{m}{d+\log(n/\delta)}\right)\cdot\left(\frac{\frac{2}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}}\omega\pi_{\min}^{2(1+\zeta^{-1})}}{5\tau} + \frac{C_1\omega^2\pi_{\min}^{2(1+\zeta^{-1})}}{\tau}\right)\max_{j\neq j'}\left\|\boldsymbol{v}_{j,j'}^t - \boldsymbol{v}_{j,j'}^\star\right\|_2^2.
\end{aligned}
\tag{SM4.10}
$$

Due to $\boldsymbol{\beta}^t \in \mathcal{N}(\boldsymbol{\beta}^\star)$ defined in (2.12), we have (SM3.11) and (SM3.12) by Lemma SM2.4.

Inserting (SM3.11) and (SM3.12) into (SM4.10) gives

$$(\kappa\rho)^{-2}A_{\text{clean}}^2 \le 1 - \frac{\pi_{\min}^{2(1+\zeta^{-1})}\omega}{\tau}\left(1 \wedge \frac{m}{d+\log(n/\delta)}\right)\left(\frac{4}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}}\left(1-\frac{2}{5}\right) + C_1\omega\,(1+4)\right)$$

$$= 1 - \frac{\pi_{\min}^{2(1+\zeta^{-1})}\omega}{\tau}\left(1 \wedge \frac{m}{d+\log(n/\delta)}\right)\left(\frac{\frac{12}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}}}{5} + 5\omega C_1\right)$$

$$\le 1 - \frac{c_0\omega\pi_{\min}^{2(1+\zeta^{-1})}}{\tau}\left(1 \wedge \frac{m}{d+\log(n/\delta)}\right), \tag{SM4.11}$$

where c_0 is the numerical constant defined in (SM3.13). We represent (SM4.11) as

$$A_{\text{clean}}^2 \le (\kappa\rho)^2\left(1 - \frac{c_0\omega\pi_{\min}^{2(1+\zeta^{-1})}}{\tau}\cdot\left(1 \wedge \frac{m}{d+\log(n/\delta)}\right)\right). \tag{SM4.12}$$

We note that by (SM3.13), c_0 is a positive absolute constant given γ and ζ. On the other hand, the choice of τ in (SM4.9) provides a bound

$$\frac{\pi_{\min}^{2(1+\zeta^{-1})}}{\tau} = \frac{\pi_{\min}^{2(1+\zeta^{-1})}}{\sqrt{\pi_{\max}} + \pi_{\min}^{1+\zeta^{-1}}} < 1.$$

Since $(1 \wedge m/(d+\log(n/\delta))) < 1$, one can set $\omega > 0$ such that $\omega c_0 < 1$, which makes the upper bound in the right-hand side of (SM4.12) a positive scalar belonging in $(0,1)$.

By following the arguments in (SM3.15) and (SM3.16), if

$$A_{\text{noise}} \le \kappa\rho\left(\frac{c_0\omega\pi_{\min}^{2(1+\zeta^{-1})}}{2\tau}\right)\left(1 \wedge \frac{m}{d+\log(n/\delta)}\right) \tag{SM4.13}$$

holds, we have

$$A_{\text{noise}} \le \kappa\rho\left(1 - \sqrt{1 - \frac{c_0\omega\pi_{\min}^{2(1+\zeta^{-1})}}{\tau}\left(1 \wedge \frac{m}{d+\log(n/\delta)}\right)}\right). \tag{SM4.14}$$

Since the upper bounds (SM4.12) and (SM4.14) satisfies (SM4.4) it suffices to show (SM4.13).

By (SM4.2), we have

$$\sqrt{\mathbb{E}_{I_t}\left\|\frac{1}{m}\sum_{i\in I_t}\nabla_{\boldsymbol{\beta}_j}\ell_i^{\text{noise}}(\boldsymbol{\beta}^t)\right\|_2^2} \lesssim \sigma\sqrt{\left(\frac{d+\log(n/\delta)}{m} \vee \frac{kd\log(n/d)+\log(1/\delta)}{n}\right)}$$

for all $j \in [k]$. After applying Jensen's inequality, we consider the choice of μ given in (SM4.8). Then, we have

$$\begin{aligned} A_{\text{noise}} &= \mu\mathbb{E}_{I_t}\left\|\frac{1}{m}\sum_{i\in I_t}\nabla_{\boldsymbol{\beta}_j}\ell_i^{\text{noise}}(\boldsymbol{\beta}^t)\right\|_2 \le \mu\sqrt{\mathbb{E}_{I_t}\left\|\frac{1}{m}\sum_{i\in I_t}\nabla_{\boldsymbol{\beta}_j}\ell_i^{\text{noise}}(\boldsymbol{\beta}^t)\right\|_2^2} \lesssim \\ &\frac{\sigma\omega\pi_{\min}^{1+\zeta^{-1}}}{\tau}\left(1 \wedge \frac{m}{d+\log(n/\delta)}\right)\sqrt{\left(\frac{d+\log(n/\delta)}{m} \vee \frac{kd\log(n/d)+\log(1/\delta)}{n}\right)}. \end{aligned} \tag{SM4.15}$$

Since (2.15) implies (SM3.18), we can choose a sufficiently large absolute constant $C > 0$ in (SM3.18) such that (SM3.18) and (SM4.15) result in (SM4.13). We complete the proof of induction argument in Step 1.

Step 2: In this step, we show that every iterate obeys

$$\begin{aligned}&\mathbb{E}_{I_t}\left\|\boldsymbol{\beta}^{t+1}-\boldsymbol{\beta}^{\star}\right\|_2 \leq \\ &\sqrt{1-\nu}\left\|\boldsymbol{\beta}^{t}-\boldsymbol{\beta}^{\star}\right\|_2 + C'\mu\sigma\sqrt{k}\cdot\left(\sqrt{\frac{d+\log(n/\delta)}{m}} \vee \sqrt{\frac{kd\log(n/d)+\log(1/\delta)}{n}}\right).\end{aligned} \tag{SM4.16}$$

In Step 1, we showed $\boldsymbol{\beta}^t \in \mathcal{N}(\boldsymbol{\beta}^\star)$. By following the argument (SM4.3), we have

$$\begin{aligned}\mathbb{E}_{I_t}\|\boldsymbol{\beta}^{t+1}-\boldsymbol{\beta}^{\star}\|_2 &\leq \mathbb{E}_{I_t}\left\|\boldsymbol{\beta}^t-\mu\frac{1}{m}\sum_{i\in I_t}\nabla_{\boldsymbol{\beta}}\ell_i^{\text{clean}}(\boldsymbol{\beta}^t)-\boldsymbol{\beta}^\star\right\|_2 + \mathbb{E}_{I_t}\left\|\frac{1}{m}\sum_{i\in I}\nabla_{\boldsymbol{\beta}}\ell_i^{\text{noise}}(\boldsymbol{\beta}^t)\right\|_2 \\ &\leq \underbrace{\sqrt{\mathbb{E}_{I_t}\left\|\boldsymbol{\beta}^t-\mu\frac{1}{m}\sum_{i\in I_t}\nabla_{\boldsymbol{\beta}}\ell_i^{\text{clean}}(\boldsymbol{\beta}^t)-\boldsymbol{\beta}^\star\right\|_2^2}}_{B_{\text{clean}}} + \underbrace{\sqrt{\mathbb{E}_{I_t}\left\|\frac{1}{m}\sum_{i\in I}\nabla_{\boldsymbol{\beta}}\ell_i^{\text{noise}}(\boldsymbol{\beta}^t)\right\|_2^2}}_{B_{\text{noise}}},\end{aligned} \tag{SM4.17}$$

where the last inequality holds by the Jensen's inequality. We first show an upper bound on B_{clean} in (SM4.17):

$$B_{\text{clean}}^2 \leq (1-\nu)\sum_{j=1}^{k}\left\|\boldsymbol{\beta}_j^t-\boldsymbol{\beta}_j^\star\right\|_2^2. \tag{SM4.18}$$

By following the argument in (SM3.23), (SM4.18) holds if there exist constants $\mu, \lambda \in (0,1)$ such that for all $\boldsymbol{\beta}^t \in \mathcal{N}(\boldsymbol{\beta}^\star)$,

$$\begin{aligned}&\sum_{j=1}^{k}\mathbb{E}_{I_t}\left\langle\frac{1}{m}\sum_{i\in I_t}\nabla_{\boldsymbol{\beta}_j}\ell_i^{\text{clean}}(\boldsymbol{\beta}^t), \boldsymbol{\beta}_j^t-\boldsymbol{\beta}_j^\star\right\rangle \\ &\qquad\geq \frac{\mu}{2}\sum_{j=1}^{k}\mathbb{E}_{I_t}\left\|\frac{1}{m}\sum_{i\in I_t}\nabla_{\boldsymbol{\beta}_j}\ell_i^{\text{clean}}(\boldsymbol{\beta}^t)\right\|_2^2 + \frac{\lambda}{2}\sum_{j=1}^{k}\|\boldsymbol{\beta}_j^t-\boldsymbol{\beta}_j^\star\|_2^2.\end{aligned} \tag{SM4.19}$$

Hence, we show (SM4.19).First, since (SM3.2) holds, (SM3.25) holds. Also, the left-hand side in (SM4.19) can be computed as (SM4.6). Thus, by (SM4.6) and (SM3.25), we obtain a lower bound on the left-hand side of (SM4.19):

$$\sum_{j=1}^{k}\mathbb{E}_{I_t}\left\langle\frac{1}{m}\sum_{i\in I_t}\nabla_{\boldsymbol{\beta}_j}\ell_i^{\text{clean}}(\boldsymbol{\beta}^t), \boldsymbol{\beta}_j^t-\boldsymbol{\beta}_j^\star\right\rangle \geq \frac{\frac{6}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}}\pi_{\min}^{1+\zeta^{-1}}}{5}\sum_{j=1}^{k}\|\boldsymbol{\beta}_j^t-\boldsymbol{\beta}_j^\star\|_2^2. \tag{SM4.20}$$

Furthermore, to obtain an upper bound on first term in the right-hand side of (SM4.19), applying (SM4.1) with the elementary inequality $\|\boldsymbol{a}+\boldsymbol{b}\|_2^2 \le 2\|\boldsymbol{a}\|_2^2 + 2\|\boldsymbol{b}\|_2^2$ provides

(SM4.21)
$$\begin{aligned}\mathbb{E}_{I_t}\left\|\frac{1}{m}\sum_{i\in I_t}\nabla_{\boldsymbol{\beta}_j}\ell_i^{\text{clean}}(\boldsymbol{\beta}^t)\right\|_2^2 &\le C_1\left(1\vee\frac{d+\log(n/\delta)}{m}\right)\Bigg(\left(\sqrt{\pi_{\max}}+\pi_{\min}^{1+\zeta^{-1}}\right)\|\boldsymbol{\beta}_j^t-\boldsymbol{\beta}_j^\star\|_2^2\\&\quad+\frac{2\pi_{\min}^{1+\zeta^{-1}}}{k}\sum_{j':j'\neq j}\left(\|\boldsymbol{\beta}_j^t-\boldsymbol{\beta}_j^\star\|_2^2+\|\boldsymbol{\beta}_{j'}^t-\boldsymbol{\beta}_{j'}^\star\|_2^2\right)\Bigg).\end{aligned}$$

Taking summation on (SM4.21) over $j\in[k]$ yields

(SM4.22)
$$\begin{aligned}&\sum_{j=1}^k\mathbb{E}_{I_t}\left\|\frac{1}{m}\sum_{i\in I_t}\nabla_{\boldsymbol{\beta}_j}\ell_i^{\text{clean}}(\boldsymbol{\beta}^t)\right\|_2^2\\&\le C_1\left(1\vee\frac{d+\log(n/\delta)}{m}\right)\left(\sqrt{\pi_{\max}}+\pi_{\min}^{1+\zeta^{-1}}+4\pi_{\min}^{1+\zeta^{-1}}\right)\sum_{j=1}^k\|\boldsymbol{\beta}_j^t-\boldsymbol{\beta}_j^\star\|_2^2.\end{aligned}$$

Putting the bounds (SM4.20) and (SM4.22) in (SM4.19) with μ chosen in (SM4.8), we have a sufficient condition for (SM4.19):

(SM4.23)
$$\frac{\frac{6}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}}\pi_{\min}^{1+\zeta^{-1}}}{5}\ge\frac{\omega\pi_{\min}^{1+\zeta^{-1}}C_1\left(\sqrt{\pi_{\max}}+5\pi_{\min}^{1+\zeta^{-1}}\right)}{2\left(\sqrt{\pi_{\max}}+\pi_{\min}^{1+\zeta^{-1}}\right)}+\frac{\lambda}{2}.$$

(SM4.23) is satisfied when we choose $\omega>0$ small enough and λ as in (SM3.29). Hence, we have shown (SM4.18) with $\nu=\mu\lambda$ where μ and λ are chosen by (SM4.8) and (SM3.29).

Next, we bound B_{noise} in (SM4.17). By (SM4.2), we obtain an upper bound on B_{noise}:

(SM4.24)
$$\begin{aligned}B_{\text{noise}}^2&=\mu^2\sum_{j=1}^k\mathbb{E}_{I_t}\left\|\frac{1}{m}\sum_{i\in I_t}\nabla_{\boldsymbol{\beta}_j}\ell_i^{\text{noise}}(\boldsymbol{\beta}^t)\right\|_2^2\\&\lesssim k\mu^2\sigma^2\left(\frac{d+\log(n/\delta)}{m}\vee\frac{kd\log(n/d)+\log(1/\delta)}{n}\right).\end{aligned}$$

Finally, putting (SM4.18) and (SM4.24) in (SM4.17) gives (SM4.16). We complete the proof of Step 2.

Step 3: We finish the proof of Theorem 3.1 using the results demonstrated in Step 1 and Step 2. By substituting the expression $\nu=\mu\lambda$, where we choose μ and λ according to (SM4.8) and (SM3.29) respectively, into (SM4.16), we obtain

$$
\begin{aligned}
&\mathbb{E}_{I_t}\|\boldsymbol{\beta}^t-\boldsymbol{\beta}^\star\|_2 \\
&(1-\mu\lambda)^{t/2}\|\boldsymbol{\beta}^0-\boldsymbol{\beta}^\star\|_2 + C_2\cdot\frac{\mu\sigma}{1-\sqrt{1-\mu\lambda}}\cdot\sqrt{k\cdot\left(\frac{d+\log(n/\delta)}{m}\vee\frac{kd\log(n/d)+\log(1/\delta)}{n}\right)} \\
&\overset{(a)}{\le}(1-\mu\lambda)^{t/2}\|\boldsymbol{\beta}^0-\boldsymbol{\beta}^\star\|_2 + C_2\cdot\frac{2\sigma}{\lambda}\cdot\sqrt{k\cdot\left(\frac{d+\log(n/\delta)}{m}\vee\frac{kd\log(n/d)+\log(1/\delta)}{n}\right)} \\
&\overset{(b)}{\le}(1-\mu\lambda)^{t/2}\|\boldsymbol{\beta}^0-\boldsymbol{\beta}^\star\|_2 + C_3\cdot\frac{\sigma}{\pi_{\max}}\cdot\sqrt{k\cdot\left(\frac{d+\log(n/\delta)}{m}\vee\frac{kd\log(n/d)+\log(1/\delta)}{n}\right)} \\
&\overset{(c)}{\le}(1-\mu\lambda)^{t/2}\|\boldsymbol{\beta}^0-\boldsymbol{\beta}^\star\|_2 + C_3\cdot\sigma k\cdot\sqrt{k\cdot\left(\frac{d+\log(n/\delta)}{m}\vee\frac{kd\log(n/d)+\log(1/\delta)}{n}\right)},
\end{aligned}
$$

where i) (a) follows from the inequality $\sqrt{1-t}<-t/2+1$ for any $t\in(0,1)$; ii) (b) holds by the choice of τ in (SM4.9); iii) (c) is a result of $\pi_{\max}^{-1}\le k$.

SM4.1. Proof of Lemma SM4.1. We will show that both (SM4.1) and (SM4.2) hold with probability at least $1-\delta/3$. Furthermore, for simplicity, we proceed on the proofs using $\boldsymbol{\beta}$ and $\boldsymbol{v}_{j,j'}$ instead of using $\boldsymbol{\beta}^t$ and $\boldsymbol{v}_{j,j'}^t$ in the statements of Lemma SM4.1. Thus, we complete the assertions in (SM4.1) and (SM4.2) by substituting $\boldsymbol{\beta}$ and $\boldsymbol{v}_{j,j'}$ with $\boldsymbol{\beta}^t$ and $\boldsymbol{v}_{j,j'}^t$ respectively.

Proof of (SM4.1): We show that with high probability, (SM4.1) holds if

$$
n \ge C_1\left(\log(k/\delta)\vee d\log(n/d)\right)k^4\pi_{\min}^{-4(1+\zeta^{-1})}, \tag{SM4.25}
$$

Note that (2.15) is a sufficient condition for (SM4.25). We proceed with the proof under the following six events, each of which holds with probability at least $1-\delta/18$. First, by the proof of (SM3.3) in Subsection SM3.1, (SM4.25) is a sufficient condition to invoke (SM3.3) with probability at least $1-\delta/18$. Next, by following the argument for (SM3.39), (SM4.25) is a sufficient condition to invoke (SM3.39) with probability at least $1-\delta/18$. Furthermore, (SM4.25) implies (SM2.13) and is a sufficient condition to invoke Lemma SM2.7 and Lemma SM2.1 with probability at least $1-\delta/18$ respectively. Hence, by following the arguments for (SM3.40), (SM3.41), and (SM3.32), (SM3.40), (SM3.41), and (SM3.32) hold with probability at least $1-\delta/18$ respectively. The last event is defined as

$$
\max_{i\in[n]}\|\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\intercal\| \lesssim d+\log(n/\delta). \tag{SM4.26}
$$

By Lemma SM1.1 and the union bound over $i\in[n]$, (SM4.26) holds with probability at least $1-\delta/18$.

Since we showed that (SM3.3), (SM3.39), (SM3.40), (SM3.41), (SM3.32), and (SM4.26) hold with probability at least $1-\delta/3$, we will move forward with the remainder of the proof by assuming those conditions are satisfied.

Let $\boldsymbol{\beta}^\star \in \mathbb{R}^{d+1}$, $\boldsymbol{\beta} \in \mathcal{N}(\boldsymbol{\beta}^\star)$, and $j \in [k]$ be arbitrarily fixed. By the argument in [SM7, Equation 7], we decompose

$$\mathbb{E}_I \left\| \frac{1}{m} \sum_{i \in I} \nabla_{\boldsymbol{\beta}_j} \ell_i^{\text{clean}}(\boldsymbol{\beta}) \right\|_2^2 = \underbrace{\frac{1}{m} \mathbb{E}_{i_1} \left\| \nabla_{\boldsymbol{\beta}_j} \ell_{i_1}^{\text{clean}}(\boldsymbol{\beta}) \right\|_2^2}_{\text{(A)}} + \underbrace{\frac{m-1}{m} \|\nabla_{\boldsymbol{\beta}_j} \ell^{\text{clean}}(\boldsymbol{\beta})\|_2^2}_{\text{(B)}}, \tag{SM4.27}$$

where we define $I := \{i_1, \ldots, i_m\} \subset [n]$ and $\nabla_{\boldsymbol{\beta}_j} \ell^{\text{clean}}(\boldsymbol{\beta})$ in (SM3.1).

Note that (SM3.3) gives an upper bound on (B):

$$\text{(B)} \lesssim \frac{m-1}{m} \left(\left(\pi_{\max} + \pi_{\min}^{2(1+\zeta^{-1})} \right) \|\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^\star\|_2^2 + \frac{\pi_{\min}^{2(1+\zeta^{-1})}}{k^2} \sum_{j' : j' \neq j} \|\boldsymbol{v}_{j,j'} - \boldsymbol{v}_{j,j'}^\star\|_2^2 \right). \tag{SM4.28}$$

It remains to show the bound on (A). By following arguments (SM3.43), we decompose $\nabla_{\boldsymbol{\beta}_j} \ell_i^{\text{clean}}(\boldsymbol{\beta})$ following

$$\nabla_{\boldsymbol{\beta}_j} \ell_i^{\text{clean}}(\boldsymbol{\beta}) = \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j\}} \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j - \boldsymbol{\beta}_j^\star \rangle \boldsymbol{\xi}_i + \sum_{j' : j' \neq j} \mathbb{1}_{\left\{\boldsymbol{x}_i \in \mathcal{C}_j \cap \mathcal{C}_{j'}^\star\right\}} \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j^\star - \boldsymbol{\beta}_{j'}^\star \rangle \boldsymbol{\xi}_i, \quad \forall i \in [n]. \tag{SM4.29}$$

Then it follows from (SM4.29) that for any $i \in [n]$,

$$\begin{aligned}
&\left\| \nabla_{\boldsymbol{\beta}_j} \ell_i^{\text{clean}}(\boldsymbol{\beta}) \right\|_2^2 \\
&\overset{\text{(i)}}{\leq} 2 \left\| \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j\}} \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j - \boldsymbol{\beta}_j^\star \rangle \boldsymbol{\xi}_i \right\|_2^2 + 2 \left\| \sum_{j' : j' \neq j} \mathbb{1}_{\left\{\boldsymbol{x}_i \in \mathcal{C}_j \cap \mathcal{C}_{j'}^\star\right\}} \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j^\star - \boldsymbol{\beta}_{j'}^\star \rangle \boldsymbol{\xi}_i \right\|_2^2 \\
&\overset{\text{(ii)}}{=} 2 \cdot \left\| \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\top \right\| \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j\}} \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j - \boldsymbol{\beta}_j^\star \rangle^2 + 2 \cdot \left\| \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\top \right\| \cdot \sum_{j' : j' \neq j} \mathbb{1}_{\left\{\boldsymbol{x}_i \in \mathcal{C}_j \cap \mathcal{C}_{j'}^\star\right\}} \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j^\star - \boldsymbol{\beta}_{j'}^\star \rangle^2 \\
&\overset{\text{(iii)}}{\lesssim} (d + \log(n/\delta)) \cdot \left(\mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j\}} \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j - \boldsymbol{\beta}_j^\star \rangle^2 + \sum_{j' : j' \neq j} \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j \cap \mathcal{C}_{j'}^\star\}} \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j^\star - \boldsymbol{\beta}_{j'}^\star \rangle^2 \right),
\end{aligned} \tag{SM4.30}$$

where (i) holds due to $\|\boldsymbol{a} + \boldsymbol{b}\|_2^2 \leq 2\|\boldsymbol{a}\|_2^2 + 2\|\boldsymbol{b}\|_2^2$; (ii) holds since $\mathcal{C}_j \cap \mathcal{C}_l^\star$ and $\mathcal{C}_j \cap \mathcal{C}_{l'}^\star$ are disjoint for any $l \neq l' \in [k]$; and (iii) holds by (SM4.26).

Applying the expectation on (SM4.30) yields

$$\begin{aligned}
&\mathbb{E}_{i_1} \left\| \nabla_{\boldsymbol{\beta}_j} \ell_{i_1}(\boldsymbol{\beta}) \right\|_2^2 \lesssim \\
&(d + \log(n/\delta)) \cdot \left(\underbrace{\frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j\}} \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j - \boldsymbol{\beta}_j^\star \rangle^2}_{\text{(a)}} + \underbrace{\frac{1}{n} \sum_{j' : j' \neq j} \sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j \cap \mathcal{C}_{j'}^\star\}} \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j^\star - \boldsymbol{\beta}_{j'}^\star \rangle^2}_{\text{(b)}} \right).
\end{aligned} \tag{SM4.31}$$

An upper bound on (b) is provided by (SM3.32). It remains to derive an upper bound on (a).

The triangle inequality provides

$$\text{(a)} \le \sum_{j'=1}^{k} \left\| \sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j \cap \mathcal{C}_{j'}^\star\}} \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\mathsf{T} \right\| \cdot \|\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^\star\|_2^2 \tag{SM4.32}$$

For the summand indexed by $j' = j$, the set inclusion, $\mathcal{C}_j \cap \mathcal{C}_j^\star \subseteq \mathcal{C}_j^\star$ yields

$$\sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j \cap \mathcal{C}_j^\star\}} \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\mathsf{T} \preceq \sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j^\star\}} \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\mathsf{T}.$$

Therefore, by (SM3.39) and (SM3.41), we have

$$\begin{aligned} \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j^\star\}} \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\mathsf{T} \right\| &\le \max_{\mathcal{I}: |\mathcal{I}| \le 2n\mathbb{P}(\boldsymbol{x} \in \mathcal{C}_j^\star)} \left\| \frac{1}{n} \sum_{i \in \mathcal{I}} \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\mathsf{T} \right\| \\ &\lesssim (\eta^2 \vee 1) \sqrt{\mathbb{P}(\boldsymbol{x} \in \mathcal{C}_j^\star)} \\ &\le (\eta^2 \vee 1) \sqrt{\pi_{\max}}, \end{aligned} \tag{SM4.33}$$

where the last inequality holds by the definition of $\pi_{\max}$. Similarly, by (SM3.40) and (SM3.41), we have

$$\left\| \sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j \cap \mathcal{C}_{j'}^\star\}} \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\mathsf{T} \right\| \lesssim (\eta^2 \vee 1) \sqrt{c} \left(\frac{\pi_{\min}^{1+\zeta^{-1}}}{k} \right), \quad \forall j' \neq j. \tag{SM4.34}$$

Then by plugging in (SM4.33) and (SM4.34) into (SM4.32), we obtain

$$\text{(a)} \lesssim \left(\sqrt{\pi_{\max}} + \pi_{\min}^{1+\zeta^{-1}} \right) \|\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^\star\|_2^2.$$

Finally, applying obtained upper bounds on (a) and (b) in (SM4.31) gives

$$\text{(A)} \lesssim \frac{(d + \log(n/\delta))}{m} \left(\left(\sqrt{\pi_{\max}} + \pi_{\min}^{(1+\zeta^{-1})} \right) \|\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^\star\|_2^2 + \frac{\pi_{\min}^{(1+\zeta^{-1})}}{k} \sum_{j': j' \neq j} \|\boldsymbol{v}_{j,j'} - \boldsymbol{v}_{j,j'}^\star\|_2^2 \right). \tag{SM4.35}$$

Putting (SM4.28) and (SM4.35) in (SM4.27) completes the proof.

Proof of (SM4.2): We proceed with the proof under the following three events, each of which holds with probability at least $1 - \delta/9$. First, (2.15) invokes (SM3.4) with probability at least $1 - \delta/9$. Next, by following the same argument in the proof of (SM4.1), (SM4.26) holds with probability at least $1 - \delta/9$. The last event is the following:

$$\frac{1}{n} \sum_{i=1}^{n} z_i^2 \le \sigma^2 \left(1 + \sqrt{\frac{C \log(1/\delta)}{n}} \right). \tag{SM4.36}$$

Since $\{z_i\}_{i=1}^n$ are i.i.d σ-sub-Gaussian random variables, the Bernstein's inequality yields that (SM4.36) holds with probability at least $1-\delta/9$.

We have shown that (SM3.4), (SM4.26), and (SM4.36) hold with probability at least $1-\delta/3$. For the remainder of the proof, we assume that those conditions are satisfied.

Then, by the argument in [SM7, Equation 7], we decompose

$$\mathbb{E}_I \left\| \frac{1}{m} \sum_{i\in I} \nabla_{\boldsymbol{\beta}_j} \ell_i^{\text{noise}}(\boldsymbol{\beta}) \right\|_2^2 = \underbrace{\frac{1}{m} \mathbb{E}_{i_1} \left\| \nabla_{\boldsymbol{\beta}_j} \ell_{i_1}^{\text{noise}}(\boldsymbol{\beta}) \right\|_2^2}_{\text{(A)}} + \underbrace{\frac{m-1}{m} \| \nabla_{\boldsymbol{\beta}_j} \ell^{\text{noise}}(\boldsymbol{\beta}) \|_2^2}_{\text{(B)}}, \tag{SM4.37}$$

where we define $I := \{i_1, \dots, i_m\} \subset [n]$ and $\nabla_{\boldsymbol{\beta}_j} \ell^{\text{noise}}(\boldsymbol{\beta})$ in (SM3.1).

(SM3.4) gives an upper bound on (B):

$$\text{(B)} \lesssim \frac{\sigma^2 kd \log(n/d) + \log(k/\delta)}{n}. \tag{SM4.38}$$

The remaining step is to obtain a bound on (A). Since we have

$$\left\| \nabla_{\boldsymbol{\beta}_j} \ell_{i_1}^{\text{noise}}(\boldsymbol{\beta}) \right\|_2^2 \le \| z_{i_1} \boldsymbol{\xi}_{i_1} \|_2^2 \le \| \boldsymbol{\xi}_{i_1} \boldsymbol{\xi}_{i_1}^\intercal \| z_{i_1}^2 \lesssim d + \log(n/\delta) z_{i_1}^2,$$

where the last inequality holds by (SM4.26), applying the expectation and (SM4.36) gives an upper bound on (A):

$$\begin{aligned} \text{(A)} \lesssim \frac{1}{n} \sum_{i=1}^n z_i^2 \left(\frac{d + \log(n/\delta)}{m} \right) &\lesssim \sigma^2 \left(1 \vee \left(\frac{\log(1/\delta)}{n} \right)^{1/2} \right) \left(\frac{d+\log(n/\delta)}{m} \right) \\ &\le \sigma^2 \left(\frac{d + \log(n/\delta)}{m} \right), \end{aligned} \tag{SM4.39}$$

where the last inequality hold by (2.15). Putting the results (SM4.38) and (SM4.39) into (SM4.37) reduces to (SM4.2).

SM5. Discussion on the proofs of [SM5, Theorem 1] and [SM4, Theorem 1]. In the proof of [SM5, Theorem 1], they claimed that $n \gtrsim \delta^{-2}$ implies [SM5, Equation (45)]. They showed that [SM5, Equation (45)] follows from [SM5, Lemmas 10 and 11]. Their [SM5, Lemma 10] presents the concentration of the supremum of an empirical measure via the VC dimension and [SM5, Lemma 11] computes an upper bound on the VC dimension of the feasible set of the maximization. According to their proof argument, the number of observations n should be proportional to the VC dimension $d \log(n/d)$ to obtain the concentration in [SM5, Equation (45)]. Their sufficient condition $n \gtrsim \delta^{-2}$ for [SM5, Equation (45)] missed the dependence on the VC dimension. We suspect that this is a typo. While it does not ruin their main result, the sample complexity in [SM5, Theorem 1] might need to be corrected accordingly. Specifically, between [SM5, Equation (32) and (33)], the parameter δ in [SM5, Lemma 6] was set to $\delta = Ck^{-2}\pi_{\min}^6$ to upper-bound the second summand in the right-hand side of [SM5, Equation (32)]. Therefore, the corrected sample complexity of [SM5, Lemma 6] increases to $\widetilde{O}(k^4 d \pi_{\min}^{-12})$ so that it dominates the sample complexity for part (b) in [SM5, Proposition 1] ($n \gtrsim kd\pi_{\min}^{-3}$). Consequently, the sample complexity in [SM5, Theorem 1] will increase by a factor $k^3\pi_{\min}^{-9}$.

Next, we report another mistake in their analysis under the generalized covariate model [SM4, Theorem 1]. They mistakenly omitted the dependence of σ in the sample complexity. A careful examination of their proof on page 48 in [SM3] will reveal that they use the same technique as in their other analysis in the Gaussian covariates case [SM5]. Therefore, we expect that their sample complexity should depend on the noise variance σ^2 to ensure that the next iterate belongs to the local neighborhood of the ground truth (refer to the proof of their Theorem 1 on page 1865 in [SM5]).

x

REFERENCES

[1] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, Learnability and the Vapnik-Chervonenkis dimension, Journal of the ACM (JACM), 36 (1989), pp. 929–965.

[2] M. Csikos, A. Kupavskii, and N. H. Mustafa, Optimal bounds on the vc-dimension, arXiv preprint arXiv:1807.07924, (2018).

[3] A. Ghosh, A. Pananjady, A. Guntuboyina, and K. Ramchandran, Max-affine regression: Provable, tractable, and near-optimal statistical estimation, arXiv preprint arXiv:1906.09255, (2019).

[4] A. Ghosh, A. Pananjady, A. Guntuboyina, and K. Ramchandran, Max-affine regression with universal parameter estimation for small-ball designs, in 2020 IEEE International Symposium on Information Theory (ISIT), IEEE, 2020, pp. 2706–2710.

[5] A. Ghosh, A. Pananjady, A. Guntuboyina, and K. Ramchandran, Max-affine regression: Parameter estimation for gaussian designs, IEEE Transactions on Information Theory, (2021).

[6] V. Koltchinskii and S. Mendelson, Bounding the smallest singular value of a random matrix without concentration, International Mathematics Research Notices, 2015 (2015), pp. 12991–13008.

[7] S. Ma, R. Bassily, and M. Belkin, The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning, arXiv preprint arXiv:1712.06559, (2017).

[8] M. Mohri, A. Rostamizadeh, and A. Talwalkar, Foundations of machine learning, MIT press, 2018.

[9] Y. S. Tan and R. Vershynin, Phase retrieval via randomized kaczmarz: theoretical guarantees, Information and Inference: A Journal of the IMA, 8 (2019), pp. 97–123.

[10] V. N. Vapnik and A. Y. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, in Measures of complexity, Springer, 2015, pp. 11–30.

[11] R. Vershynin, High-dimensional probability: An introduction with applications in data science, vol. 47, Cambridge university press, 2018.